

Trust and how to sustain it

Against a background of declining public trust in traditional institutions, scientists must work to retain their high public-confidence ratings. There are warnings and lessons to be learned from the events of 2002.

If opinion polls are to be believed, scientists enjoy great public respect and trust. In an August Harris poll, Americans ranked 'scientist' as the most prestigious profession, edging out 'doctor', which had topped the list for the previous five years. In another Harris poll taken in November, 68% of respondents said they trust scientists to tell the truth — more than the number who trusted the president, and fifth behind teachers, doctors, professors and police officers.

But public trust in all professions, from judges to journalists, has dropped noticeably since a similar survey taken last year. And we don't need pollsters to tell us that faith in institutions, from the stock market to churches, has recently been under strain. In such times, scientists would do well not to take their favourable rating for granted. Things could easily take a turn for the worse. Two of the growth industries for scientists are biotechnology and defence, both of which are politically charged, and in the context of the war on terrorism, are now related. A protracted war in Iraq or a steep economic decline could erode public respect for scientists if they are identified with failing policies.

Another risk comes from pushing, or appearing to push, controversial practices opposed by large segments of the population, such as genetic manipulation of foods or stem-cell research. This is more than a question of who's 'right'. Too many scientists dismiss opposition based on emotion or differences in world-view, arrogantly believing that if the public is simply told the facts, they will fall in line. Researchers who work for profit-making ventures should be particularly careful, as their motives are more open to suspicion. In a poll taken in March for the British Royal Society, more than half of the respondents thought that science funding is becoming too commercialized, and said they wanted more influence over the research agenda. Scientists will keep the public's trust as long as they consider citizenship as well as scholarship.

A third threat comes from recent scandals, such as that involving physicist Jan Hendrik Schön's falsification of data (see page 728).

Scientific misconduct is rare and damages the image rather than the long-term substance of science. So far, these seem like isolated cases of misconduct, and the labs involved have reacted quickly and responsibly. But the Catholic church's more serious scandals can serve as a cautionary tale. Last year, 90% of people quizzed in a Harris poll said that they trusted the clergy to tell the truth; this has now fallen to 64%. The public expects scientists, like priests, to have high standards. Not that researchers are saints or sages. But science is meant to serve truth and the advancement of knowledge above self-protection or profit. Putting self-interest first is precisely how the church got itself into trouble. Scientists and their institutions need to react promptly to misconduct, even though the burdens of doing so are usually onerous.

Trust in science could also be diminished by people who exploit scientific uncertainty for political ends, such as by casting doubt on the evidence for global warming or evolution. A few 'sceptics' appearing on TV can confuse a public that expects monolithic truth from science. All that scientists can do is explain that scientific 'truth' keeps changing in the light of new evidence, and provide the bigger picture.

Scientists are constantly exhorted to do a better job of explaining, and sometimes selling, what they do to the public. And well they should, as long as the public-relations campaigns stay honest and don't veer into deceit. It may not be the researchers themselves who do the deception, but the organizations they work for, such as biotech companies that tout gene therapy as the answer to all of life's problems.

In societies where market values increasingly trump all others, where advertising and spin are pervasive, and where broadcast news — the main source of information for most people — is increasingly partisan, the public has fewer places to turn to for objective, neutral information. Science can be of great service to society here. Out of more than just self-interest, then, scientists should be careful not to trade away the public trust, or allow others to trade it away for them. ■

Managing CERN's industries

Europe's huge particle-physics laboratory is now in need of both brilliant science and outstanding management.

Following the appointment of Robert Aymar as the new director general of the European particle-physics laboratory CERN (see page 721), the laboratory has a good chance of re-establishing the full confidence of its member states following recent budgetary problems. Aymar is known for taking as much pleasure from the good management of physics as from the physics itself. And good management will be paramount between now and 2007, as the world's next major accelerator, the Large Hadron Collider (LHC), is constructed and launched. Magnets are being tested, the gigantic detectors are being constructed, and the massive computational analysis based on Grid supercomputer networks is being developed. Through all of these projects, it will be essential to maintain confidence in what is, in effect, industrial production of the key LHC components.

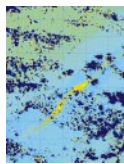
The closure of the Large Electron-Positron collider (LEP) in 2000 was the toughest and most controversial decision made by the current director general, Luciano Maiani, given the apparent signatures of the

long-sought Higgs boson from the LEP detectors. Subsequent analysis has shown those signatures to be less suggestive than was originally thought. The Large Hadron Collider will probably either reveal the Higgs for what it is or plunge high-energy physics into a conceptual crisis if it can't be found.

Despite the absence of a large accelerator, CERN does not have to wait until 2007 to accomplish strong physics. Its smaller facilities have recently created antihydrogen, definitively measured CP violation — a still mysterious asymmetry in particle decays — and observed a crucial state of nuclear matter in which the quarks and gluons that make up neutrons and protons become 'deconfined'.

To continue these programmes and install the LHC will require tough decisions as obstacles inevitably arise. Aymar will be tested to his limits, but cannot do his job single-handedly. Others in CERN's team will have to earn their laurels not only by excellent experimentation, but also by first-rate project management. ■

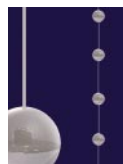




Iron rule
Plan for carbon
sequestration
left all at sea
p722



Rocky rocket
Explosion throws
future of Ariane 5
into doubt
p723



Small print
Academy report
shows support for
neutrino projects
p724



Prickly issue
Scottish hedgehogs
face cull to protect
local birds
p727

Stanford announces private plan to get stem-cell research moving

Erika Check, Washington

Stanford University in California is joining a growing list of US universities that are establishing privately funded stem-cell institutes.

In announcing its Institute for Cancer/Stem Cell Biology and Medicine on 10 December, Stanford became the latest institution to attempt to use non-federal money to try to resolve funding and oversight dilemmas in the controversial field of research on human embryonic stem cells. The facility will be headed by Irving Weissman, a prominent stem-cell researchers at Stanford.

The Stanford scientists plan to do basic research on stem cells, and develop new therapies for chronic diseases and cancer. Philip Pizzo, dean of Stanford University's Medical School, said it was "quite feasible" that they would also use a method called therapeutic cloning, or somatic-cell nuclear transfer, to create new embryonic stem-cell lines.

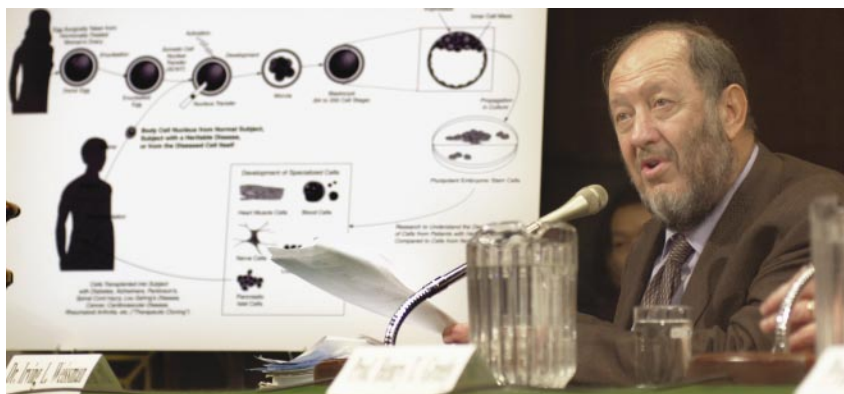
US government regulations expressly prohibit the use of federal money to do this — raising the thorny question of how free universities such as Stanford are to do it, when they depend so heavily on federal government support for other activities.

Two other US universities — the University of California, San Francisco, and Johns Hopkins University in Baltimore, Maryland — opened similar institutes this year.

President George Bush declared in August 2001 that publicly funded researchers would be allowed to work only on stem-cell lines that had already been derived from embryos. And some in the US Congress would like to ban the production of embryonic stem cells altogether, whether researchers are publicly funded or not.

Several states, including California, have passed laws supporting the procedure — partly in an attempt to attract biotechnology money. And patient-advocacy groups have convinced some key senators to oppose further restrictions on the research.

Meanwhile, scientists are pushing their universities to find creative ways to fund the work, which could lead to profitable patents and treatments. A handful of institutions have taken the plunge — but others have remained reticent.



Irving Weissman will head the Stanford institute, which may create new embryonic stem-cell lines.

"Doug Melton [a Harvard University biologist] and I have spent countless hours in the past two years trying to stimulate Harvard and the Boston area into doing a coordinated stem-cell programme," says George Daley of the Whitehead Institute for Biomedical Research in Cambridge, Massachusetts. "But I think they're going to be slow to come to the idea."

This is largely because universities are keen to avoid open conflict with the federal government. Thomas Murray, president of

the Hastings Center, an independent bioethics institute in Garrison, New York, says that such caution is well founded.

"It would not surprise me if the Bush administration attempted to halt or create as many difficulties as possible for researchers in California or Maryland or any other place supportive of the work," Murray says. "And if you want to make a hard time for somebody and you have the resources of the federal government, you can be creative." ■

CERN collider project back on track

Alison Abbott, Munich

The governing council of CERN, the European particle-physics laboratory, has approved an eight-year financial plan that will focus tightly on its main project, the Large Hadron Collider (LHC).

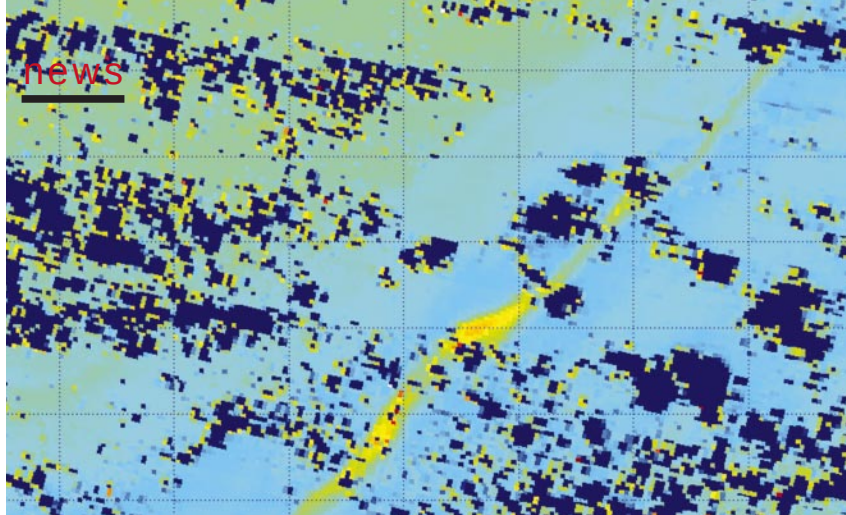
The plan's approval should end a period of uncertainty at the lab, after the emergence of hefty cost overruns on the LHC, and problems with the manufacture of its 1,000 or so superconducting magnets (see *Nature* 413, 441; 2001). The project is now due to be finished in 2007, two years later than planned.

"There have been hiccups and bumps with the LHC, as happens with very big projects," says Luciano Maiani, CERN's director general. "But although resources

are tight, we are now on course to make it happen."

The council also elected Robert Aymar, a French plasma physicist, to take over from Maiani in 2004. Aymar is currently director of ITER, an international project to build an experimental fusion reactor. But he is well acquainted with the LHC, having chaired the external review committee established last year to advise the CERN council on how to bring the project back under control.

CERN's previous heads have all been high-energy physicists, but Aymar says he plans to make an asset of his different background. "I will be free of any prejudice," he says, adding that he will appoint a deputy who is more familiar with CERN's scientific work. ■



Sea change: a satellite image of a 200-km algal bloom (orange) caused by iron fertilization.

Ocean tests raise doubts over use of algae as carbon sink

Rex Dalton, San Francisco

Plans to use marine microorganisms as a sink for atmospheric carbon dioxide appear to have suffered a setback. Early results of an experiment designed to test the concept indicate that the process is significantly less effective than originally thought — and may also have undesirable side effects.

If confirmed, the results, which are being prepared for submission to a number of journals, would weaken the argument that oceanic carbon sequestration can act as a significant component in controlling the build-up of greenhouse gases in the atmosphere.

“The experiment was a tremendous success at every level,” says Kenneth Coale of the Moss Landing Marine Laboratories near Monterey, California, who led the team of researchers, “but some of the results give us cause for concern.”

The team was investigating the effectiveness of iron fertilization, which involves adding iron sulphate to the ocean surface to encourage the growth of phytoplankton. These single-celled algae absorb carbon dioxide from the ocean, which in turn causes more of the gas to migrate from the atmosphere into the water. The huge algal blooms caused by the iron fertilization may then be eaten by other organisms or sink into the deep ocean, taking the carbon with them.

Coale and his team spent 28 days studying patches of iron sulphate spread on the ocean south of New Zealand. Known as the Southern Ocean Iron Fertilization Experiment (SOFEX), the project was the most comprehensive effort yet to study oceanic iron fertilization.

During the experiment, three research vessels dodged icebergs and ugly weather in a region chosen for its significant impact on natural sequestration of carbon dioxide from the atmosphere. Two of the research

vessels — the *Roger Revelle* and the *Melville* — repeatedly spread an iron sulphate solution on the ocean surface. As a result, the team recorded a bloom of phytoplankton 200 kilometres long.

The researchers revealed their preliminary findings at a meeting of the American Geophysical Union (AGU) in San Francisco on 9–10 December. Although the bulk of the carbon stayed near the surface, Ken Buesseler, an oceanographer at Woods Hole Oceanographic Institution in Massachusetts, said that the experiment showed for the first time that carbon can be transported below 100 metres by iron fertilization.

But the process is not as efficient as predicted. The team’s analysis indicates that one tonne of iron spread at the ocean surface could force 1,000 tonnes of carbon below 100 metres, Buesseler told the AGU meeting. Previous laboratory experiments had estimated that one tonne of iron would sink 100,000 tonnes of carbon. More carbon may sink over a longer period, he suggested, but longer experiments are needed to determine how much. “I’m not certain the oceans can ever solve our carbon dioxide problem,” says Coale.

Oliver Wingenter, a chemist at New Mexico Tech in Socorro, reported on the activity of other gases produced by phytoplankton. Wingenter found that emissions of methyl bromide, which can deplete the protective ozone layer at high altitude, increased sharply. And emissions of isoprene — a hydrocarbon that generates greenhouse gases at low altitude — rose “dramatically”, he says.

Kenneth Johnson, an oceanographer from Monterey Bay Aquarium Research Institute who was chief scientist on the *Roger Revelle*, says that these unintended consequences “aren’t great” for the feasibility of using the technique on a large scale. ■

Mass resignation forces Italian rethink on university funds

Alison Abbott, Munich

A tax on smoking could deflect a funding crisis that set off widespread protests at Italian universities last week — including the dramatic mass resignation of the leaders of all 72 Italian public universities.

The university rectors handed their resignations to the minister of universities and research after the Italian parliament’s chamber of deputies approved a 3% cut in the universities’ budget for next year. The rectors had been lobbying for an increase in overall budget to cover the increased costs of salaries, which are fixed centrally.

“We were being pushed to breaking point,” says Piero Tosi, rector of the University of Siena and head of the Italian University Rectors’ Conference. “We decided to resign because we were no longer able to assure our essential services for teaching and research.”

Italian students have expressed concern over the declining competitiveness of their universities, and possible increase in fees, in a series of demonstrations and occupations of university buildings.

Many professors are particularly worried about the consequences of the budget squeeze for research, as most universities’ funds are tied up in salaries, with little left over to support core activities. “Cutting the budget would not affect the lives of those professors who do nothing anyway,” says Dario Braga, a chemist at the University of Bologna, “but would hurt only the dynamic members of the faculty who want to do research.”

But the universities also attracted criticism from some academics, who say they should sort out their management problems before asking for more money. The universities were given autonomy in 1994 after decades of centralized administration. Since then, complains Francesco Giavazzi, an economist at Bocconi University in Milan, they have expanded unwisely. “These organizational issues need to be fixed, otherwise extra money would only go to waste,” he says.

Italian finance minister Giulio Tremonti says that he hopes not only to maintain the university budget at the 2002 level, but also to create a €200-million (US\$200-million) research fund to be spent at the discretion of prime minister Silvio Berlusconi. The measures may well persuade the rectors to withdraw their resignations, but are unlikely to please Italy’s smokers — Tremonti plans to fund his proposal with a 10-cent tax on a packet of cigarettes. ■

ESA weighs up future after Ariane explosion

David Adam, London

It was intended to be a show of strength, but after the spectacular failure of the new 'heavy-lifter' Ariane rocket, the only thing hoisted into space by the European Space Agency (ESA) last week was a weighty doubt over its future in the launch business.

Faced with a grim record of 4 failures from just 14 Ariane 5 launches, some observers were starting to ask if Europe's 25-year-old rocket programme is more trouble than it's worth. But most space scientists appear confident that ESA can ride out the latest failure and go ahead with launches of most of its planned scientific satellites.

An investigation is already under way into what caused the souped-up €150-million (US\$150-million) Ariane 5 ECA to veer dangerously off course and self-destruct just three minutes into its maiden flight from Kourou, French Guiana, on 11 December. One of the investigation's first tasks is to determine whether the fault has any bearing on the classic Ariane 5 design — which is supposed to take the Rosetta satellite mission on the first leg of its journey to rendezvous with a comet in just a few weeks' time.

"We have to understand what happened first, but it's most probable that Rosetta will launch on schedule on 12 January," Jean-Yves Le Gall, director of the ill-fated mission, told reporters at a press conference in Kourou.

The spacecraft is intended to rendezvous

with the comet Wirtanen in 2011, when it will place one probe in orbit and land another on the nucleus (see *Nature* **417**, 889; 2002). The mission aims to observe the comet as it heats up on its approach to the Sun and begins to spew out gas and dust. But if it is to keep its appointment it must be launched before the end of next month.

The rocket's loss is a heavy blow for both ESA and Arianespace, the company it set up to develop and operate the launchers. Each is banking on Ariane 5 ECA's success to win contracts to launch commercial satellites.

"This really couldn't have come at a worse time for Ariane," says Patrick French, an analyst with consultants Frost and Sullivan in Mountain View, California, who specializes in the satellite industry. The market in commercial satellite launching is depressed because of the slump in the telecommunications industry, French points out. With no military and government contracts, delays could force ESA to request extra cash from its member states to keep the Ariane programme afloat.

The money will probably be found, French says: "Europe is not going to abandon Ariane because it will still want independent access to space."

David Southwood, ESA's director of science, insists that the European space industry will "rise again" and still has a bright future. "You don't get into this business



Dark skies: Europe's Ariane 5 ECA 'heavy lifter' blew up just three minutes after launch.

J. AMET/AFP

without realizing that failures happen," he says. "It's a setback in the short term, but I don't doubt that a launcher of that capability will emerge."

Although developed primarily for the commercial market, a new class of heavy-lifter rockets will also be used to launch the next generation of deep-space research probes, such as the planned James Webb space telescope, Southwood says. ■

Australian bushfire centre rises from the flames

Carina Dennis, Sydney

With bushfires raging around Sydney for the second consecutive summer, Australia is turning to science for guidance on land and fire management. On 10 December, science minister Peter McGauran announced that A\$25 million (US\$14 million) will be spent on a multisite centre devoted to the study of bushfires.

The Bushfire Cooperative Research Centre (CRC) will commence work next July, and will involve participants from several universities, fire and emergency services, national parks and land-management groups.

Announced as part of a package of 12 new centres of this type, the bushfire initiative will be funded initially for seven years with the federal money and a further A\$87 million allocated over that period by its various participants.

"The centre will have a focus on reducing risk to life and property," says Leon Collett, programme coordinator of the Australasian Fire Authorities Council, a representative

body for fire and emergency services and land-management agencies that led the bid for the centre.

The CRC will have its headquarters in Melbourne, but the research will be carried out across the sites of the participating groups. "The history of bushfire research in Australia has been too small, unfocused and not well integrated. What the CRC does is assemble a critical mass for scientists working on a broad set of problems," says Collett.

The centre will investigate fire behaviour and ecology, smoke hazards, building design and construction materials, and will develop education programmes for the public. It will also investigate the impact of prescribed burning — the contentious practice of lighting controlled fires to remove hazardous build-ups of vegetation.

But some researchers have expressed disappointment at the centre's emphasis on fire management near urban areas — an issue of pressing relevance to

Australia's heavily populated southeastern states. They point out that most bushfires actually occur in the less-populated northern reaches of the continent. "We would have preferred a more national approach to bushfire research," says Jeremy Russell-Smith, a fire-management consultant at the Bushfires Council of the Northern Territory.

This was the third attempt by various parties to set up a bushfire research centre. The success was due to having "the right mix" of participants, says Rodney Weber, a mathematician at a campus of the University of New South Wales in Canberra who plans to model bushfire ecology for the CRC.

This year's fires have already wreaked significant destruction around Sydney, with about 730,000 hectares burned in New South Wales so far, signalling another season of damage and uncertainty around Australia's largest city. The fires are being exacerbated by searing temperatures, strong winds and drought conditions. ■

Neutrino projects are poles apart, says panel

Geoff Brumfiel, Washington

It's not just distance that separates the two neutrino projects planned by the United States — their scientific objectives are clearly different too, according to a review by the National Academy of Sciences.

The report was commissioned earlier this year by the White House Office of Science and Technology Policy (OSTP) to clarify the degree of overlap between the proposed IceCube neutrino detector in Antarctica and the National Underground Science Laboratory (NUSL), which would probably be built at Lead, South Dakota (see *Nature* **417**, 5; 2002).

The academy panel was chaired by Barry Barish, a physicist at the California Institute of Technology in Pasadena. This month's report describes the two projects as distinct, saying that there is "essentially no overlap or redundancy in their primary science goals".

John Marburger, director of the OSTP, says that although the report confirms the projects are scientifically distinct, they will end up competing for money in any case.

Neutrinos are subatomic particles with no electric charge and very small mass. They seldom interact with other particles and are not deflected by electromagnetic fields, so they can travel unhindered through



Simulation: strings of light-sensitive detectors in IceCube, one of two proposed neutrino projects.

turbulent regions of space. "Recent discoveries have created special opportunities to use neutrinos in new ways to advance our knowledge of the Universe and the laws that govern it," the panel finds.

The planned projects both aim to detect these elusive particles, but from different sources and for different reasons. IceCube would monitor a cubic-kilometre block of Antarctic ice for the rare flashes of light that occur when neutrinos interact with water

molecules. The sheer size of IceCube means that it should detect neutrinos generated by powerful events outside our Galaxy, giving astrophysicists a new way to observe phenomena such as black holes. The neutrinos might also shed light on other high-energy particles that inhabit intergalactic space.

By contrast, the NUSL, a laboratory that would have facilities 2–4.5 km below ground, would study the properties of neutrinos generated closer to home — by the Sun and by particle accelerators on Earth. Its experiments would be aimed at improving researchers' understanding of neutrinos themselves and of the fundamental nature of matter.

But despite their differences, the two projects have some things in common: each is projected to cost about a quarter of a billion dollars to build, and each has already lined up some enthusiastic backers in Congress. IceCube, for example, secured \$15 million in funding in 2002 and could receive another \$25 million next year (see *Nature* **418**, 573; 2002).

But the panel's sterling endorsement of both projects is no guarantee that either will proceed. The National Science Foundation and Department of Energy will now review the findings and decide whether to include them in future budget requests. ■

DARWIN RIANTO & BILL BELLO

Corporate chiefs told to follow animal urges

John Whitfield, London

Three British zoologists have set up a consultancy firm that plans to sell ideas gleaned from the study of animal behaviour to corporate clients.

Animals and businesses face many of the same problems, argues one of the group, Alex Kacelnik of the University of Oxford. "A petroleum company forages for oil in much the same way as a starling might forage for worms," he says. The basic choice

they face, he points out, is whether to stay in a field that is productive but in decline, or to move on and take a chance elsewhere.

Kacelnik's work on starlings has led him to believe that humans and other animals approach risk in similar ways. He now plans to apply his models of how animals cope with uncertainty through the consultancy firm Oxford Risk Research and Analysis (ORRA), which he founded last month with two other Oxford zoologists, John Krebs, chair of the UK Food Standards Agency, and Ed Mitchell. ORRA aims to study companies' decision-making processes, identify areas where risks are being misjudged, and recommend remedies.

The consultancy has already started work with clients in the pharmaceutical and energy industries, a university statement said. In work for one multinational client, it found that decision-makers tended to be too cautious about doing business in unfamiliar regions, but were too gung-ho about investing where they knew. ORRA recommended changes in where and when the client made decisions, and called for more contact between its different operations.

It is too early to say whether any of ORRA's clients will follow its recommendations, Kacelnik says. The company has worked on three projects so far, and aims to have about a dozen, along with five full-time employees, by the end of its first year's operations.

Outside observers say that the consultancy's approach may prove alluring. "Companies are willing to explore any viable approach to dealing with risk and using it to their advantage," says Richard Apostolik, president of the Global Association of Risk Professionals, based in Jersey City, New Jersey. But to prosper, he adds, ORRA will have to show clients tangible results.

Animal behaviour can teach us much about human decision-making, says psychologist Daniel Read of the London School of Economics. But he urges caution: "Animals have the product of millions of years of evolution, but they can't take that knowledge and apply it to new environments." In situations where humans get the chance to stop and think things through, Read argues, animal behaviour may not be a very useful model. ■



Complaints prompt inquiry into Indian plagiarism allegations

New Delhi A letter of complaint from Stanford University physicists has led to an inquiry being established into alleged plagiarism by Balwant Singh Rajput, a high-energy physicist and vice-chancellor of Kumaun University.

Rajput and his group are accused of taking almost 30 papers by different authors, making minor changes and then publishing them under their own names. The governor of Uttaranchal, the northern Indian state in which Kumaun University is based, said last week that a three-member panel will look into the allegations. The move comes after seven Stanford physicists wrote to the president of India, A. P. J. Abdul Kalam, asking him to act on the alleged plagiarism.

Rajput says he is confident that the inquiry will find no truth in the allegations. He told *Nature* that members of his group had put his name on papers without his consent.

Researchers free to sift wheat genome data

Washington The chemicals and food company DuPont has unexpectedly made public much of its proprietary information on the wheat

genome, one of the largest and most complex genomes in the plant world.

The company, based in Wilmington, Delaware, last week deposited 200,000 wheat expressed-sequence tags (ESTs) with the gene database GenBank. ESTs — small pieces of gene sequences — are valuable tools for geneticists because they can be used to identify expressed genes. The deposit doubles GenBank's wheat genome data.

Jim Miller of DuPont's Crop Genetics Research and Development division says that the donation will facilitate collaborations among crop scientists. It has been welcomed by the International Triticeae Mapping Initiative, an academic collaboration that is mapping the genomes of crops such as wheat and barley.



Dust the job: DuPont's work on the wheat genome could bring benefits to crop research.

Biotech firm gets licence for stem cell

Washington Technology used to create a promising type of adult stem cell has been licensed to Athersys, a biotechnology company in Cleveland, Ohio.

The cells, which seem to be able to generate a range of mature cells, were isolated from human bone marrow by Catherine Verfaillie's group at the University of Minnesota in Minneapolis (Y. Jiang *et al. Nature* **418**, 41–49; 2002). Verfaillie says that the Athersys agreement will make the cells available to academic researchers without onerous restrictions, although the company will retain the option to license any discovery made using them.

The ability of adult stem cells to generate different tissues has been hotly debated, but Verfaillie says that other groups are doing similar experiments, and will soon replicate her results.

Missile defence body under attack for redefining failure

Washington The government agency charged with developing a system for intercepting warheads aimed at the United States is artificially inflating the success rate of its tests, according to a report from the Union of

Concerned Scientists (UCS), a Boston-based science watchdog group.

UCS physicist Lisbeth Gronlund says the problem lies with the Missile Defense Agency's (MDA) definition of failure. The MDA has reclassified some failures from over a decade ago as successes, and has indiscriminately lumped together the tests of different missile-defence systems, the report says.

The report also accuses the MDA of unfairly defining success in terms of the final seconds of intercepts. On 11 December, for example, the MDA announced that it "was not able to complete" a test because the interceptor failed to separate from its booster. "That kind of statement is sort of fair," says Gronlund, "but the interceptor would have to separate from the booster in the real world."

Medics sick of rising conference costs

London The free exchange of ideas at medical conferences is under threat from the growing practice of charging delegates to attend specific sessions, a medical journal asserts.

An editorial in this month's *The Lancet Oncology* (3, 707; 2002) says that conference organizers risk changing the "culture and spirit" of conferences by turning them into "pick-and-mix" affairs at which attendees

choose which sessions to go to in advance.

The practice could also lead to researchers with limited funds having to ration their attendance. Tickets costing up to US\$100 were needed for some sessions at this year's annual meeting of the American Society of Clinical Oncology in April in Orlando, Florida. The journal says that a society member attending one such session a day would pay an "astonishing" total of US\$675 in conference fees — more than twice the advertised US\$275.

Likewise, some sessions at the European Society for Medical Oncology (ESMO) meeting in Nice, France, in October cost an additional E20 (US\$20) each. "We believe that our pricing policy is appropriate and fair," says an ESMO spokesman, pointing out that attendance at the annual congress has steadily grown over the past 10 years.

Calls for cull are no yolk for egg-hungry hedgehogs

London Christmas is coming, and in Scotland's western isles it's not only the turkeys that are getting nervous. Thousands of hedgehogs on the islands of North Uist, South Uist and Benbecula are facing a bleak midwinter after scientists recommended they be culled to protect bird populations.

Introduced to the isles in 1974, the

hedgehog community has soared to over 5,000. And with a taste for birds' eggs, their arrival has coincided with a catastrophic collapse in populations of ground-breeding waders, including ringed plover and dunlin.

A planned hedgehog cull was postponed in the summer by conservation group Scottish Natural Heritage (SNH) following public protests. But it now seems the most likely option, as alternatives such as moving the animals to the mainland or keeping them in captivity appear to have been ruled out. The SNH is expected to decide the fate of the hedgehogs this week.



At the sharp end: hedgehogs on Scottish islands may be killed for eating birds' eggs.

SCOTTISH NATURAL HERITAGE

Controversy has been the watchword in a year dogged by dispute. Misconduct revelations, clashes over transgenic crops, and confusion over stem cells have threatened to overshadow triumphs in fields from palaeoanthropology to fundamental physics. *Nature's* reporters recount the year's talking points.

Misconduct

The stars who fell to Earth

He was, according to the report that brought his career to an abrupt end, “a hard working and productive scientist”; a bright and personable young man whose stunning work on molecular-scale electronic devices had him marked by some as a future Nobel prizewinner.

But today, we know Jan Hendrik Schön as the perpetrator of the most outrageous fraud ever to tarnish the physical sciences, fabricating or falsifying data in at least 16 papers, many published in *Nature* and *Science*. “At the minimum,” concluded the report that sealed his fate, published in September¹, “Hendrik Schön showed reckless disregard for the sanctity of data in the value system of science.”

For physicists, who were accustomed to viewing such aberrations as the preserve of the biological sciences, the conclusions of the expert committee assembled by Schön's employer, Lucent Technologies' Bell Laboratories in Murray Hill, New Jersey, made for chilling reading. And unfortunately, Schön couldn't be dismissed as a one-off. A few months earlier, the Lawrence Berkeley National Laboratory in California fired physicist Victor Ninov for allegedly fabricating data in work of similarly high profile: the claimed creation of elements 116 and 118, the heaviest known². Both scientists have acknowledged making some errors, but neither has confessed to the damning lists of alleged misdeeds compiled against them.

Some research leaders have lauded the way in which the two labs have handled the cases, arguing that Schön and Ninov's precipitous falls from grace show that science is self-correcting. “In the end, nature is the checker,” agrees Pier Oddone, deputy director of the

Berkeley lab, who handled the Ninov probe. “Experiments have to be reproducible.”

But for others, the cases confirm that the cancer of fraud now confronts all scientific disciplines, raising questions about the responsibility of co-authors and stressing the importance of training in standards of scientific conduct. Those who have studied misconduct say that miscreants often start with minor infractions, building up to widespread falsification of data should their earlier misdeeds go undetected. And experiences in biology — where the issues are, sadly, nothing new — raise doubts about the research community's ability or desire to confront such spiralling mendacity.

Twin tracks

The careers of Schön and the Bulgarian-born Ninov followed eerily similar trajectories. Both trained in Germany, at the University of Constance and the Heavy Ion Research Center (GSI) in Darmstadt, respectively. Both left for the bright lights of leading US labs, only to be exposed by the harsh searchlights of misconduct investigations.

In Schön's case, both the quality and quantity of his work marked him out. He seemed to be a wizard in testing the properties of electronic devices that he had fashioned from thin films of organic molecules, and in inducing superconductivity in the spherical carbon molecules known as ‘buckyballs’. During the two years preceding the launch of the investigation into his work, Schön churned out a staggering 80 research papers.

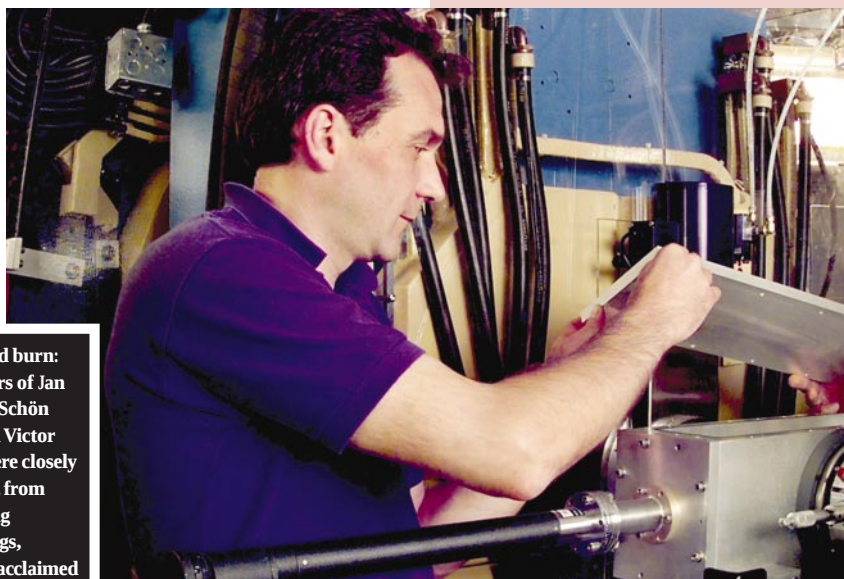
Although, in retrospect, this sounds suspicious, such prodigious output was not seen as impossible for a scientist who seemed so dedi-

cated. But tellingly, the investigating committee, headed by applied physicist Malcolm Beasley of Stanford University in California, concluded: “None of the most significant physical results was witnessed by any co-author.” The report cleared Schön's colleagues of misconduct, but argued that the case raised wider questions about the responsibility of co-authors, for which “no clear, widely accepted standards of behaviour exist”.

Ninov's field of heavy-element research did not allow for such prolific output, but his alleged crimes against science were no less audacious. The case also illustrates the emerging pathology of misconduct: his employers concluded that, after starting with small instances of data fabrication that did not change an experiment's main conclusions, Ninov based his ‘discovery’ of elements 116 and 118 on a web of falsified data.

Ninov's co-authors withdrew the claimed discovery of these elements in July 2001, having failed to replicate the results. Shortly afterwards, his former colleagues at the GSI were re-analysing data representing the signals of elements 111 and 112 (refs 3, 4). Data had clearly been altered, says GSI physicist Sigurd Hofmann, and computer records indicated that the changes were made from Ninov's account. Hofmann recalls thinking: “Is this a joke?” If so, it was a very bad one, and Ninov was unable to offer an explanation.

The data involved didn't make a significant difference to outcome of those experiments. Did this represent a trial run made by Ninov to see whether he could get away with falsification? Or did he find himself sucked into an addictive pattern of escalating misconduct? Whatever the explanation, in July this year Hofmann and his colleagues reported on their repeat of the experiments on elements 111 and 112, confirming that the earlier papers' main conclusions are still valid in the absence of Ninov's “spuriously created” data⁵.



Crash and burn: the careers of Jan Hendrik Schön (left) and Victor Ninov were closely matched, from promising beginnings, through acclaimed publications, to ultimate disgrace and discredit.

But that wasn't the case for the paper on elements 116 and 118, which was formally retracted during the same month⁶. By then, Ninov's career at the Lawrence Berkeley lab had unravelled, after his colleagues delved into data files and discovered the same type of manipulation identified at the GSI.

These two cases have prompted reforms. In November, for instance, the American Physical Society issued guidelines on ethics and the responsibilities of co-authors, intended to prevent a repeat of the situation in which Schön could falsify data without anyone looking over his shoulder. The guidelines list "co-authors who are accountable for the integrity of critical data reported in the paper, carry out the analysis, write the manuscript, present major findings at conferences, or provide scientific leadership" as bearing responsibility for all of a paper's contents.

Rules or role models?

But some physicists have already complained that the new guidelines place too great a burden of responsibility on the supervisors of junior researchers. And the US Office of Research Integrity, which reviews allegations of scientific misconduct relating to projects funded by the National Institutes of Health (NIH), has been trying unsuccessfully for several years to require all NIH grant holders to receive ethics training. The proposal has languished in the face of opposition from the Federation of American Societies for Experimental Biology (FASEB). "My position is that you learn by watching role models, not by being in a mandated or instructed programme," says FASEB president Steven Teitelbaum, a pathologist at Washington University in St Louis, Missouri.

The German system that trained Schön and Ninov, on the other hand, had earlier introduced codes of scientific good practice

and new mechanisms for investigating allegations of misconduct in the wake of the case of Fried-

helm Herrmann and Marion Brach — cancer researchers who systematically fabricated data in scores of publications while at the Max Delbrück Centre for Molecular Medicine in Berlin in the early 1990s.

While the scientific community wrestles with how to prevent similar scandals in the future, the immediate repercussions of the Schön and Ninov cases are still being felt. Ninov is fighting to clear his name, and has hired an attorney who specializes in employment disputes. In January, the University of California, which manages the Berkeley lab, will hold a grievance hearing on Ninov's firing.

Science has already published a brief note retracting eight of Schön's papers⁷. *Nature* will shortly run separate retractions for each of Schön's papers; in the meantime, the online versions carry a warning about the conclusions of the Beasley report. Other journals are going through similar processes.

Hopefully, these will run their course more rapidly than in previous misconduct cases. In July, *Nature* investigated the status of the 94 papers authored by Herrmann and Brach that were listed two years previously in an official inquiry as "definitely or probably" containing manipulated data. Many had not been retracted, and some journal editors were not even aware of the misconduct investigation⁸.

Science may be self-correcting, but sometimes it is a painfully slow process. ■

Rex Dalton

1. http://www.lucent.com/news_events/researchreview.html
2. Ninov, V. et al. *Phys. Rev. Lett.* **83**, 1104–1107 (1999).
3. Hofmann, S. et al. *Z. Phys. A* **350**, 277–280 (1995).
4. Hofmann, S. et al. *Z. Phys. A* **354**, 229–230 (1996).
5. Hofmann, S. et al. *Eur. Phys. J. A* **14**, 147–157 (2002).
6. Ninov, V. et al. *Phys. Rev. Lett.* **89**, 039901(E) (2002).
7. Bao, Z. et al. *Science* **298**, 961 (2002).
8. Abbott, A. & Schwarz, J. *Nature* **418**, 113 (2002).

Highlight: Genomics

Sorry, dogs — man's got a new best friend

Compared with the media frenzy that greeted the publication of the human genome, coverage of the draft mouse sequence, unveiled this month¹, was a relatively sedate affair. But for many scientists, the mouse genome deserves at least equal billing, as it provides the key to unlock the secrets of our own DNA.

The two genomes, it turns out, are remarkably similar: 99% of mouse genes have a direct human counterpart. It is not a unique set of genes that make us human rather than murine, but rather the way that they are regulated.

On one level, this humbling similarity exposes the daunting complexity of mammalian biology. But it also means that the mouse — newly named as man's best friend — is here to help promote our self-knowledge, and to spur medical advances.

Armies of mutant mice are being created to study human diseases, and to understand myriad aspects of our biology. Molecular tools to interpret sequences are also being developed apace. For instance, a new set of mouse complementary DNA clones² will make it easier to investigate gene function by generating 'knockout' mice.

What next? Another team has already posted on the Internet a preliminary assembly of the rat genome³, and a peer-reviewed publication should follow next year. This is good news for physiologists, for whom the rat is the favourite experimental animal.

Alison Abbott

1. Mouse Genome Sequencing Consortium *Nature* **420**, 520–562 (2002).
2. The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I and II Team *Nature* **420**, 563–573 (2002).
3. www.hgsc.bcm.tmc.edu





Food for thought: evidence that Mexico's diverse maize (left) is contaminated by transgenic strains is hotly contested; GM proponents argue that the technology will benefit those in rice-growing areas.



Agribiotech

More heat than light

Another year, another controversy: that was the story in the perennially contentious area of genetically modified (GM) crops.

In 2002, arguments centred on David Quist and Ignacio Chapela's study of Mexican maize. It was simultaneously a bitter ideological feud among biologists at a single US university and a flashpoint between the agribiotech industry and anti-GM activists over the acceptability of transgenic crops in the developing world — which is becoming a key battleground. And it illustrated how, in this field, the quest for scientific truth is conducted in a minefield of opinion and accusations of vested interests.

The story began in November 2001, when Chapela, an assistant professor at the University of California, Berkeley, and Quist, his postdoc, published a paper in *Nature*¹ reporting that a 'promoter' sequence from transgenic crops was present in native Mexican maize, and had fragmented throughout the genome. It was a provocative finding, as Mexico is the world's centre of genetic diversity for maize, and operates a moratorium on commercial GM planting.

Supporters of GM technology pored over the paper, and soon argued that the pair's results were an artefact of the molecular techniques they had used^{2,3}. Quist and Chapela disagreed, but the further evidence they produced⁴ failed to convince all of the experts, and in April *Nature* published the exchange with an editorial note⁵ saying that, in hindsight, the original paper's publication was unjustified.

By this time, pro- and anti-GM websites were buzzing with claims and counter-claims,

and journalists were realizing that many of Quist and Chapela's scientific opponents also had Berkeley connections. Indeed, some critics had clashed with Chapela and Quist over the pair's opposition to Berkeley's controversial deal with Syngenta, which gives the Swiss-based agribiotech firm privileged access to the findings of the university's plant scientists. It then emerged that some Internet postings attacking Quist and Chapela had been made from computers at a public-relations firm retained by GM giant Monsanto of St Louis, Missouri. Clearly, this was not solely a technical debate.

The scientific facts remain unclear. For months, Exequiel Ezcurra, president of the National Institute of Ecology in Mexico City, has been suggesting that Mexican scientists have replicated Quist and Chapela's findings, but the results have yet to appear in a peer-reviewed journal. Meanwhile, scientists at CIMMYT, the International Maize and Wheat Improvement Center in Texcoco, Mexico, have drawn a blank in their search for transgenic DNA in Mexican maize.

Why is everyone so agitated about the alleged contamination? In part, the answer is that developing countries such as Mexico now represent the front line in the war over GM technology. Agribiotech companies have largely saturated the North American market, and face a bleak future in Europe thanks to consumer opposition and the imminent introduction of strict labelling for GM food. The consumers and farmers of Central and South America, Asia and Africa represent the firms' main potential for growth.

Some companies are keen to stress the benefits that the technology could bring to

rice farmers, for example. This year saw the publication of drafts of the entire rice genome^{6,7}, plus

finished versions of two of its 12 chromosomes^{8,9}, and GM proponents are full of ideas for how the crop could be improved. Among the most appealing is the prospect of launching a second 'Green revolution' by radically overhauling the efficiency with which rice makes sugars by photosynthesis¹⁰.

But such goals remain distant. In the eyes of many activists, GM crops are primarily tools to advance the profits of agribiotech firms and wrest economic control of the food chain from small-scale farmers. This helps to explain one of the year's most perplexing developments: the decision of several southern African countries, though facing famine, to reject US offers of food aid containing GM grain. Most have since relented, provided that the grain is milled to prevent planting. But as *Nature* went to press, Zambia was still holding out.

Elsewhere in the developing world, attitudes are diverging. India emerged as a GM proponent in March when it approved commercial plantings of cotton engineered by Monsanto to produce bacterial insecticide. In Brazil, meanwhile, the official line seems set to swing against transgenic agriculture after the victory of Luiz Inacio Lula da Silva, a left-winger allied to the country's small-scale farmers, in October's presidential election. Lula's administration replaces a government that approved commercial planting of GM soya, only to be blocked by a legal challenge from Greenpeace and a local consumer group. The case is still awaiting resolution.

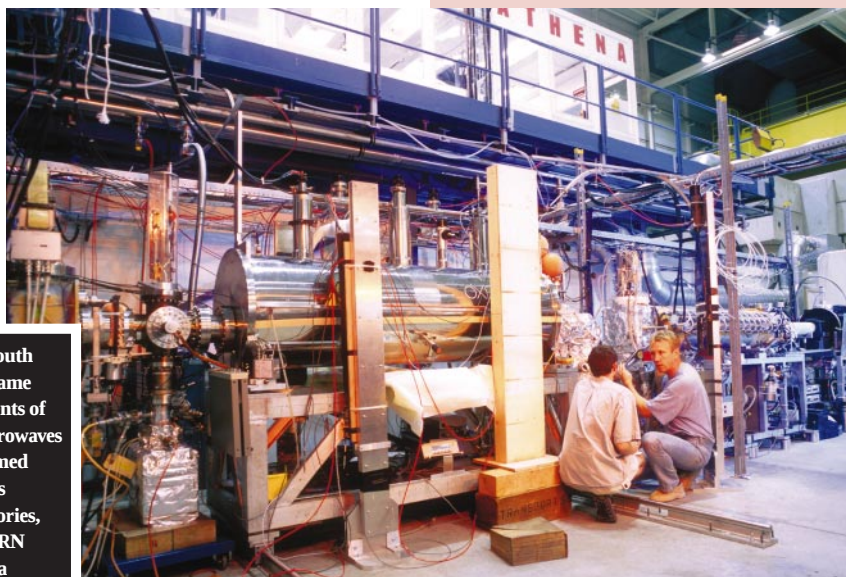
Doubts remain, however, about the developing world's ability to implement and monitor policies on GM agriculture. Despite the *de facto* moratorium imposed by Brazil's court proceedings, much of the soya grown in the southern state of Rio Grande do Sul is



From the South Pole (left) came measurements of cosmic microwaves that confirmed cosmology's leading theories, while at CERN near Geneva (right), two teams captured atoms of antihydrogen.

derived from GM grain imported illegally from Argentina. And unapproved GM cotton varieties have reportedly been widely planted in India, hampering attempts to monitor the environmental and economic impact of the officially sanctioned crops.

Peter Aldhous



1. Quist, D. & Chapela, I. H. *Nature* **414**, 541–543 (2001).
2. Metz, M. & Fütterer, J. *Nature* **416**, 600–601 (2002).
3. Kaplinsky, N. et al. *Nature* **416**, 601–602 (2002).
4. Quist, D. & Chapela, I. H. *Nature* **416**, 602 (2002).
5. *Nature* **416**, 600 (2002).
6. Yu, J. et al. *Science* **296**, 79–92 (2002).
7. Goff, S. A. et al. *Science* **296**, 92–100 (2002).
8. Sasaki, T. et al. *Nature* **420**, 312–316 (2002).
9. Feng, Q. et al. *Nature* **420**, 316–320 (2002).
10. Surridge, C. *Nature* **416**, 576–578 (2002).

Highlight: Antihydrogen

Holding up a mirror to physics' world view

Atoms from the mirror world of antimatter were captured and analysed for the first time this year. Two teams at CERN, the European Laboratory for Particle Physics near Geneva, have created large numbers of long-lived antihydrogen atoms, which can be used to test fundamental theories about the Universe.

In total, the teams produced thousands of antihydrogen atoms by using magnetic fields to bring together antiprotons and anti-electrons. This was a considerable technical achievement in itself, but the real interest lies in studying the properties of antihydrogen. Theory suggests that it should mirror the properties of hydrogen, but no one knows for sure, says Rolf Landua, spokesman for the ATHENA collaboration, which announced its results in September.

The standard model of fundamental particles and forces holds that hydrogen and antihydrogen should have the same properties and obey the same rules, but it can't predict why the Universe is almost devoid of antimatter. Finding a difference between matter and antimatter could lead to an explanation, and perhaps force physicists to reformulate the standard model.

A second CERN team, called ATRAP, has also captured antihydrogen atoms², and has subsequently made preliminary measurements of their most excited energy states³. "There's still a lot of work to do," says Gerald Gabrielse of Harvard University, spokesman for ATRAP.

"This is just the start," agrees Landua. "Everything is fresh; we haven't exploited all of our potential."

Geoff Brumfield

1. Amoretti, M. et al. *Nature* **419**, 456–459 (2002).
2. Gabrielse, G. et al. *Phys. Rev. Lett.* **89**, 213401 (2002).
3. Gabrielse, G. et al. *Phys. Rev. Lett.* **89**, 233401 (2002).

Cosmology

It all adds up

The embers of the Universe's primordial fire continue to provide cosmologists with data against which to test their ideas. And the most reassuring news this year has been that their favourite theories seem up to the job.

September saw the release of new results on the cosmic microwave background (CMB), the blanket of microwave radiation that pervades the Universe. The microwave photons that make up the CMB date from just 300,000 years after the Big Bang. By analysing enough of them, cosmologists can detect faint records of conditions in the youthful Universe.

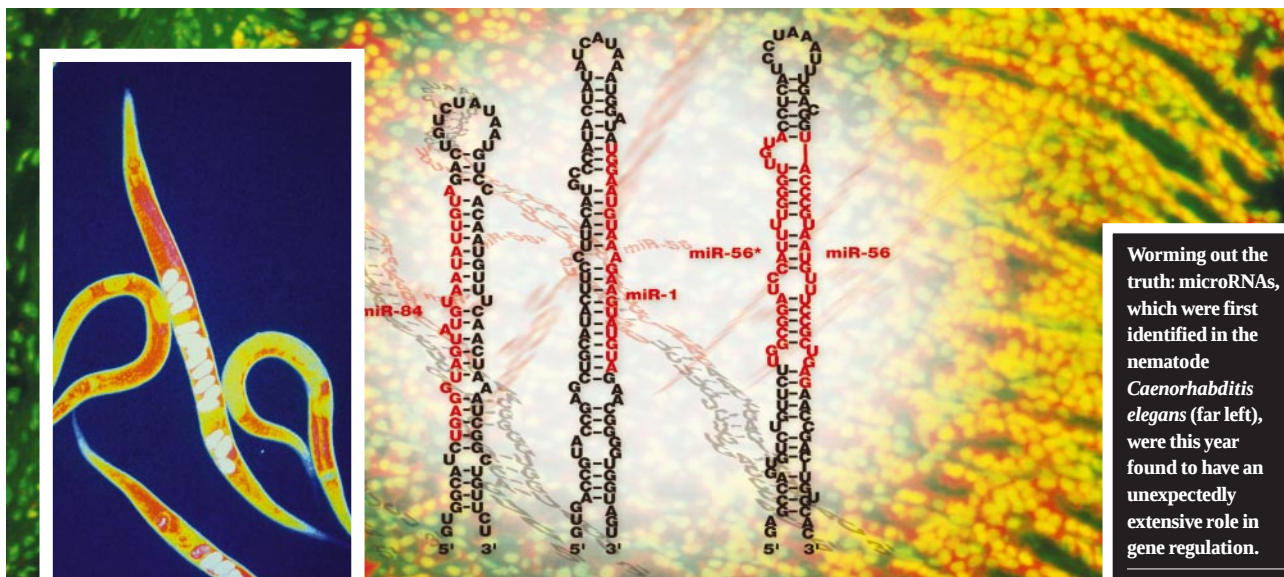
For 271 days over the past two years, researchers used the Degree Angular Scale Interferometer (DASI), an observatory at the US Amundsen-Scott South Pole Station, to study microwave photons coming from two small patches of the Antarctic winter sky. The images from each day were combined into a single super-high-resolution picture, which revealed a slight polarization in the microwave background.

The polarization was created by electrons surfing the waves of energy that swept through the early Universe, says John Carlstrom, an astronomer at the University of Chicago and leader of the DASI project, who

announced the news at the International Workshop on Particle Physics and the Early Universe, held in Chicago. The results are published in this issue of *Nature*^{1–3}. As electrons sped along the front of the wave, they reflected photons in a preferential direction, causing the radiation in certain areas of the Universe to become polarized. When the electrons eventually fell into orbit around protons, the radiation escaped, taking with it a record of the energy waves.

The polarization data back up other results to come from studies of the CMB. During the early 1990s, temperature differences in the CMB were detected⁴. Cosmologists suspect the differences were caused by clumps of matter in the early Universe, which seeded the growth of the web of galaxies we see today. According to the leading theory, these clumps of matter were created by the energy waves — so detecting evidence of the waves was important. "The polarization is a unique signature," says Carlstrom. "If it wasn't there, we'd have to throw out this theory of waves, which means we'd have to throw out all of our recent interpretations of the CMB."

NASA's Microwave Anisotropy Probe satellite, launched in June 2001, will provide a more detailed all-sky picture of the polar-



ization sometime early next year. But the polarization may retain a secret. The energy waves are thought to have been triggered by a rapid enlargement of the early Universe, known as inflation. This would also have created gravity waves, which would lead to tiny whirlpools in the CMB polarization. Detecting these whirlpools would provide the best evidence yet for inflation. Measuring

such vortices will require more sensitive instruments, however. "No one knows for sure exactly how to do it — yet," says Lyman Page, a cosmologist at Princeton University in New Jersey. ■

Geoff Brumfiel

1. Leitch, E. M. *et al. Nature* 420, 763–771 (2002).
2. Kovac, J. M. *et al. Nature* 420, 772–787 (2002).
3. Zaldarriaga, M. *Nature* 420, 747–748 (2002).
4. Smoot, G. *et al. Astrophys. J. Lett.* 396, L1–L4 (1992).

Small RNAs

The genome's guiding hand?

They can be thought of as biology's 'dark matter': tiny RNAs that don't encode any protein. But 2002 has seen an avalanche of discoveries about their roles in influencing gene activity — lending some credence to the radical idea that small RNAs hover 'above' the genome, providing a matrix of regulatory control.

MicroRNAs (miRNAs) are about 22 nucleotides long, and were first identified in the nematode *Caenorhabditis elegans*^{1,2}. In *C. elegans*, they regulate the activity of specific genes by binding to messenger RNAs (mRNAs) and preventing their translation to proteins.

Until recently, many experts thought that such examples were interesting anomalies. But they are now realizing that miRNAs are involved in gene regulation in a wide variety of organisms, and researchers are finding links between miRNAs and the phenomenon of RNA interference (RNAi). The latter mechanism, which is exploited by biologists for gene-silencing studies, is thought to be a natural defence against invading viruses. It uses an enzyme called Dicer to cut double-stranded viral RNA into small interfering RNAs (siRNAs), again about 22 nucleotides long. These then bind to other viral RNA,

targeting it for destruction.

Dicer also creates miRNAs, by cutting them from longer, hairpin-shaped RNAs transcribed from the genome, but that was where the similarity was thought to stop. In August, however, it emerged that miRNAs can also function in the RNAi pathway and cause their target mRNAs to be degraded, if they perfectly match its target sequence³.

Speculation was mounting that miRNAs represented an unexplored layer of gene regulation but, without knowing the identity of their targets, their role remained unclear. This again changed in the summer, when the mRNA targets of miRNAs were identified in the weed *Arabidopsis thaliana* — most of those studied seem to be involved in early plant development^{4,5}. This was remarkable progress, given that miRNAs were only identified in plants a few months earlier^{6,7}.

At the same time, new functions for small RNAs were being found. Tiny RNAs now appear to do a whole lot more than just target mRNAs. They seem to be associated with various 'epigenetic' phenomena — the inheritance of features that do not involve genetic sequence changes. For instance, a startling connection has been made between

small RNAs and the silencing of gene activity in tightly packed regions of the genome: deleting genes that encode components of the RNAi pathway led to a loss of gene silencing in the fission yeast *Schizosaccharomyces pombe*^{8,9}.

Small RNAs also seem to be capable of reshaping the genome. The ciliated protozoan *Tetrahymena thermophila* possesses two nuclei, the larger of which loses roughly 15% of its DNA during the cell's development — and this seems to be guided by small RNAs¹⁰.

Finally, some researchers are taking a lead from the presumed natural function of RNAi by exploring the use of small RNAs as anti-viral agents. siRNAs targeted at viral genes can inhibit the replication of HIV-1 or poliovirus in cultured cells^{11–13}. What's more, siRNAs targeted at the host receptors used by HIV to enter the cell can also block infection¹³.

Of course, results in cultured cells don't always translate to the clinic. "The challenge will be in the delivery," says Andy Fire of the Carnegie Institution of Washington in Baltimore, co-discoverer of RNAi in *C. elegans*. But there are early signs that siRNAs targeted at human hepatitis C virus can function when injected into mice¹⁴. These small nucleotide strings could yet be big news for medicine. ■

Carina Dennis

1. Lee, R. C., Feinbaum, R. L. & Ambros, V. *Cell* 75, 843–854 (1993).
2. Reinhart, B. J. *et al. Nature* 403, 901–906 (2000).
3. Hutvagner, G. & Zamore, P. A. *Science* 297, 2056–2060 (2002).
4. Llave, C., Xie, Z., Kasschau, K. D. & Carrington, J. C. *Science* 297, 2053–2056 (2002).
5. Rhoades, M. W. *et al. Cell* 110, 513–520 (2002).
6. Llave, C., Kasschau, K. D., Rector, M. A. & Carrington, J. C. *Plant Cell* 14, 1605–1619 (2002).
7. Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B. & Bartel, D. P. *Genes Dev.* 16, 1616–1626 (2002).
8. Volpe, T. A. *et al. Science* 297, 1833–1837 (2002).
9. Hall, I. M. *et al. Science* 297, 2232–2237 (2002).
10. Mochizuki, K., Fine, N. A., Fujisawa, T. & Gorovsky, M. A. *Cell* 110, 689–699 (2002).
11. Gitlin, L., Karelsky, S. & Andino, R. *Nature* 418, 430–434 (2002).
12. Jacque, J.-M., Triques, K. & Stevenson, M. *Nature* 418, 435–438 (2002).
13. Novina, C. D. *et al. Nature Med.* 8, 681–686 (2002).
14. McCaffrey, A. P. *et al. Nature* 418, 38–39 (2002).



Fine words: despite a hostile reception for US Secretary of State Colin Powell at the sustainable-development summit (right), parallel meetings have opened useful dialogue to address the needs of the developing world.



Sustainable development

Now we're talking

This summer's World Summit on Sustainable Development in Johannesburg was an easy target for critics. With governments reluctant to commit more money and US President George W. Bush deciding not to attend, the event looked set to be an expensive failure.

Politically speaking, this pessimism was warranted. Although timescales were agreed for restoring fisheries and improving access to clean water and sanitation, none was legally binding. But researchers and scientific organizations interested in sustainable development made strides both during and after the meeting. An assessment of how science can contribute to sustainable development was launched at the summit, and plans for new research centres and fresh sources of funding have been unveiled since.

For scientists, the important discussions at Johannesburg took place in a series of parallel meetings convened by the Paris-based International Council for Science (ICSU) — an umbrella group of national and international scientific organizations. The organizers were pleasantly surprised by the reaction of those who attended. "There was a strong signal from the international scientific community that it was focused on the goal of sustainable development," says Pamela Matson, a biogeochemist at Stanford University in California who studies the consequences of land-use decisions in the developing world.

Emboldened by the response, ICSU announced in October that it plans to establish centres in Asia and the Pacific, Africa, Latin America and the Caribbean, and the Arab region to promote the best science for sustainable development. "They will draw on

traditional knowledge and scientific expertise to address local problems," says ICSU president-elect Goverdhan Mehta, director of the Indian Institute of Science in Bangalore. The centres will also identify the key threats to sustainable development in each region.

India has already agreed to provide office space and cover some of the costs for one centre, and other potential hosts are currently being courted. If successful, the centres should help to overcome some of the key obstacles identified at Johannesburg, such as the need for sustainable-development policies and practice to be tailored to specific cultures, ecosystems and economies — and for them to be developed within the region for which they are intended.

Other scientific bodies have also launched new initiatives since the summit. The US National Academies plan to hold regular meetings to promote sustainable development among US researchers, and have raised a \$10-million budget for these gatherings. First on the agenda for the inaugural meeting, to be held in Washington in January, is a discussion of how scientists can be encouraged to think about sustainable-development research suggested by non-governmental organizations and the private sector. "We need to aggressively move away out of the

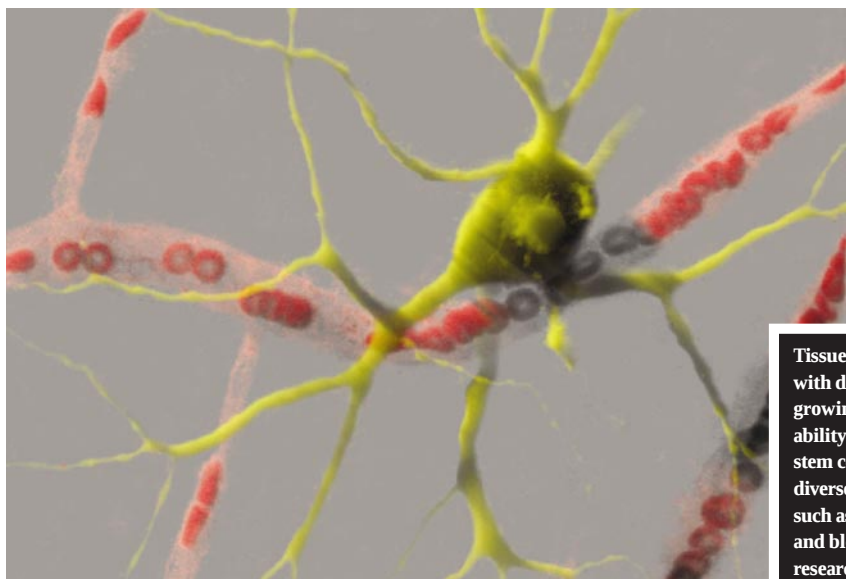
normal mode of doing studies at the request of government," says William Clark, an expert on sustainable development at Harvard University who is involved in the academies' initiative.

November saw the launch of the Sixth Framework Programme, the European Union's latest five-year research funding plan. Of the programme's E17.5-billion (US\$17.6-billion) budget, E1.85 billion is earmarked for sustainable-development research; E600 million of that will be spent in the developing world, compared with just E1.4 million under the fifth Framework.

The US government, frequently criticized by activists for its lack of involvement in sustainable development, has also taken some positive steps, in part because of concerns over the link between poverty and national security. In October, for example, the CIA published a report¹ noting that HIV may decimate the young adult population in many countries, destabilizing armed forces, the business elite and social institutions. Failed nations tend to become breeding grounds for conflict and terrorism, which explains why September's annual review of US security policy² emphasized the role of sustainable development. Nevertheless, many experts feel that the political 'hawks' who dominate the US administration may limit the extent to which initiatives to promote sustainability can complement traditional military and intelligence approaches to national security.

The enormous effort needed to convert scientific advances into solutions to improve the lot of the world's poorest people was highlighted in October, with the publication of the genome sequences of the deadliest malaria parasite, *Plasmodium falciparum*³, and its mosquito vector, *Anopheles gambiae*⁴. New approaches to fighting malaria will eventually follow, but the celebration of these landmarks was tempered by the knowledge that poverty

Johannesburg was a beacon of hope — it got scientists to talk about the right things.



Tissue trauma:
with doubts
growing over the
ability of adult
stem cells to create
diverse tissues
such as nerves
and blood (left),
research on banks
of embryonic
stem cells (right)
remains a priority.

denies existing treatments and preventative measures to many millions of people.

The road ahead won't be easy. But Johannesburg provided a beacon of hope because it got scientists to talk about the right things, suggests Clark. "The very good news is that we got a summit where development and environment were on the agenda," he says.

"We now have an opportunity that we didn't have before." ■

Tom Clarke

1. Gordon, D. F. *The next wave of HIV/AIDS: Nigeria, Ethiopia, Russia, India, and China* (National Intelligence Council report ICA 2002-04D, Washington DC, 2002).
2. <http://www.whitehouse.gov/nsc/nss.html>
3. Gardner, M. J. et al. *Nature* **419**, 485–511 (2002).
4. Holt, R. A. et al. *Science* **298**, 129–149 (2002).

heart can repair damage caused by a heart attack⁹. In an acrimonious showdown in

February, at a meeting on cardiovascular biology in Keystone, Colorado, Chuck Murry of the University of Washington in Seattle revealed that he had failed to find a single new heart-muscle cell after similar experiments.

But advocates of adult stem cells remain undeterred. In November, Stanford's Helen Blau fought back with new evidence that, after injury at least, adult stem cells do differentiate in previously unexpected ways. When muscles purged of their own stem cells were damaged by exercise, stem cells from bone marrow came to the rescue¹⁰. The exact cells responsible remain to be identified, Blau notes, and may be different from the blood stem cells studied by others. And a high-profile study by Catherine Verfaillie and her colleagues at the University of Minnesota, Minneapolis, had earlier convinced most researchers that, at least under specific culture conditions, adult bone-marrow stem cells can become capable of contributing to almost any other tissue type¹¹.

With the potential of adult stem cells still unclear, researchers are agreed that continued work on embryonic stem (ES) cells, which definitely can produce a wide range of tissue types, is crucial. And the therapeutic potential of ES cells was reinforced over the year in a series of encouraging studies^{12–14}. This is heightening the frustration felt by researchers in countries — including the United States and Germany — where government funding is restricted to a few approved ES-cell lines. President Bush approved the use of 64 lines in 2001, but only five are currently being shipped, and some of those grow poorly or have undergone successive rounds of

Stem cells

Articles of faith adulterated

Believing in miracles takes conviction — and 2002 saw researchers' faith in the healing power of adult stem cells sorely tested, as the cells' regenerative abilities became the focus of intense debate.

At issue was some scientists' belief that adult stem cells — long thought to spawn only a limited range of cell types, such as blood cells — can 'transdifferentiate' and generate other tissues such as nerves and skin. Their conviction comes largely from reports that stem cells plucked from bone marrow and transplanted into mice contribute to various tissues throughout the body^{1–3}. Such cells could help to treat conditions from Parkinson's to heart disease, they hope.

But such studies — particularly those by newcomers to the field — have come under increasing attack. In March this year, sceptics seized on two reports suggesting that, rather than switching fates, transplanted cells might masquerade as muscle or nerve cells by fusing with existing cells, creating potentially sickly tissues. Groups led by Austin Smith of the University of Edinburgh, UK, and Naohiro Terada of the University of Florida in Gainesville, showed that adult stem cells

from bone marrow or brain fuse with embryonic stem cells when the two are grown in the same dish^{4,5}. Similar fusions by other types of stem cell might explain the results from transdifferentiation experiments, sceptics argued.

This sent advocates of transdifferentiation scurrying to check whether their new cells might also be a consequence of fusion. Many researchers now accept that, at least in muscle and liver, where cell fusion occurs naturally, this phenomenon might explain some of their results. "I wouldn't be surprised if we find fusion in muscle," says Diane Krause of Yale University in New Haven, Connecticut, "but so far we haven't."

Meanwhile, a series of publications disputed earlier findings about the regenerative powers of bone-marrow stem cells^{6–8}. In one report, Irving Weissman of Stanford University in California injected just one of these cells into mice and traced its progeny with green fluorescent protein. Apart from blood cells, it spawned only eight cells of any other tissue⁹.

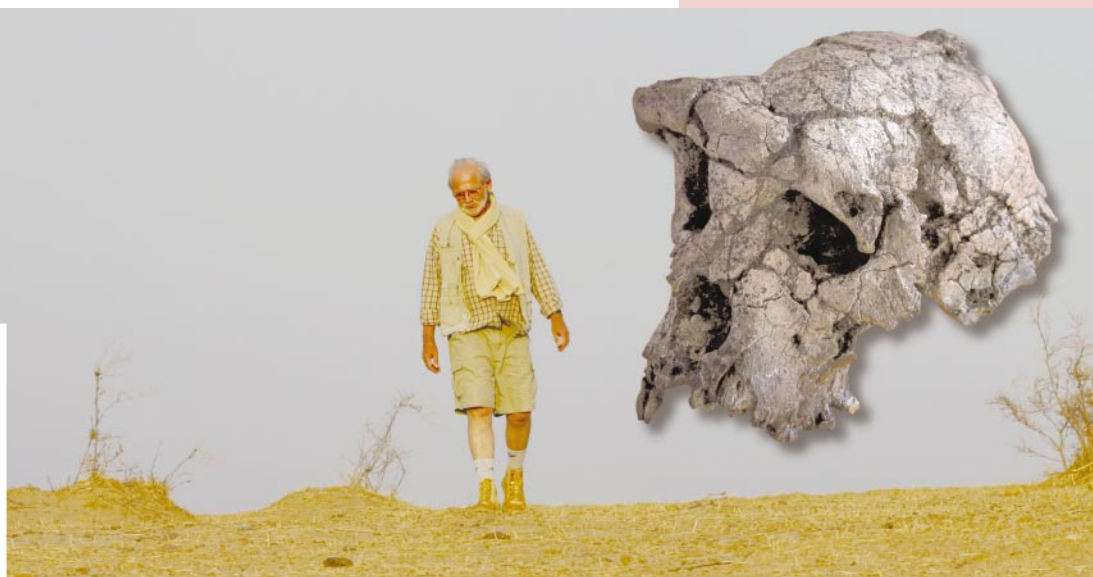
Problems of reproducibility also plagued another celebrated study, which claimed that blood stem cells pumped into a mouse

Head start for humanity: Michel Brunet's discovery in Chad of the 6- to 7-million-year-old skull of Toumai has thrown open the question of where and when our hominid ancestors first arose.

culturing — which might have altered their characteristics.

Other countries, including Britain, Singapore and Israel, are happy to let researchers isolate fresh ES-cell lines, even from cloned embryos. Elsewhere, such as in China, national regulations are still being prepared. But with the Republican party, which includes many pro-lifers in its ranks, taking over the Senate, pressure for a US ban on the production of cloned human embryos for stem-cell research will escalate. Both scientifically and politically, the climate for researchers in the field remains turbulent.

Helen Pearson



1. Brazelton, T. R., Rossi, F. M. V., Keshet, G. I. & Blau, H. M. *Science* **290**, 1775–1779 (2000).
2. Mezey, É., Chandross, K. J., Harta, G., Maki, R. A. & McKercher, S. R. *Science* **290**, 1779–1782 (2000).
3. Krause, D. S. *et al. Cell* **105**, 369–377 (2001).
4. Terada, N. *et al. Nature* **416**, 542–545 (2002).
5. Ying, Q.-L., Nichols, J., Evans, E. P. & Smith, A. G. *Nature* **416**, 545–548 (2002).
6. Morshead, C. M., Benveniste, P., Iscove, N. N. & van der Kooy, D. *Nature Med.* **8**, 268–273 (2002).
7. Castro, R. F. *et al. Science* **297**, 1299 (2002).
8. Wagers, A. J., Sherwood, R. I., Christensen, J. L. & Weissman, I. L. *Science* **297**, 2256–2259 (2002).
9. Orlic, D. *et al. Nature* **410**, 701–705 (2001).
10. LaBarge, M. A. & Blau, H. M. *Cell* **111**, 589–601 (2002).
11. Jiang, Y. *et al. Nature* **418**, 41–49 (2002).
12. Rideout, W. M., Hochedlinger, K., Kyba, M., Daley, G. Q. & Jaenisch, R. *Cell* **109**, 17–27 (2002).
13. Kim, J.-H. *et al. Nature* **418**, 50–56 (2002).
14. Wichterle, H., Lieberam, I., Porter, J. A. & Jessell, T. M. *Cell* **110**, 385–397 (2002).

Setback: Gene therapy

Shining hopes dented — but not dashed

For gene therapists, the news in October that a child in a gene-therapy trial for severe combined immunodeficiency disease (SCID) had developed leukaemia was a devastating blow. The SCID trial, led by Alain Fischer at the Necker Hospital for Sick Children in Paris, was the field's most successful, having cured nine children of this life-threatening illness.

Fischer is now trying to figure out the cause of the tragedy, and whether it could happen again. It seems that his patient's cancer started when the retroviral vector that carried a corrective gene into the boy's body activated a gene called *LMO-2*, causing one of his cells to proliferate uncontrollably. The patient is now responding to treatment.

Gene therapists are quick to stress the differences between the current setback and the 1999 death of Jesse Gelsinger in a gene-therapy trial at the University of Pennsylvania in Philadelphia. Gelsinger was not expected to be cured by the treatment that killed him, and faults were later found in the management of the trial.

Although trials in several countries were put on hold after Fischer's announcement, a consensus seems to be emerging that they should resume — albeit with revised procedures for informed consent and improved monitoring of patients.

Germany, which halted all trials involving retroviral vectors in the light of earlier data from animal experiments, has already given the green light to some of its affected trials, and advisory committees to the US Food and Drug Administration and the National Institutes of Health have recommended restarting SCID trials in the United States.

Erika Check

Palaeoanthropology

Face to face with our past

The President of Chad calls him Toumai, meaning 'hope of life'. His discoverers, led by palaeontologist Michel Brunet of the University of Poitiers in France, named him *Sahelanthropus tchadensis*, declaring him to be the oldest known human ancestor, or hominid¹. But a handful of naysayers argue that the specimen is a forerunner of modern apes, and say that it should be named *Sahelpithecus*.

Whatever you call him, the skull that graced the cover of the 11 July issue of *Nature* is one of the most talked about discoveries in the history of palaeoanthropology. Most experts were awestruck, and pleased that Brunet's long career had finally been so richly rewarded. "Fieldwork has to be broadly conceived and pursued over a long period," says Clark Howell of the University of California, Berkeley. "Brunet's discovery is a wonderful example."

The first shock was the fossil's age, which was determined from the geology of its isolated desert location near Lake Chad, as well as accompanying vertebrate fossils, to be between 6 million and 7 million years². The location itself was a surprise, indicating that early humans were evolving much farther west in Africa than was previously thought. If Brunet's assertion that Toumai is a hominid is correct, East Africa's Rift Valley can no longer be seen as the exclusive cradle of humanity.

Toumai could turn many cherished ideas about human evolution on their heads, as some of his features seem to be more advanced than those of presumed human ancestors that lived much later. Some experts now believe that he represents the tip of an iceberg of hominid diversity that existed more than 5 million years ago.

But this is palaeoanthropology, a field in



Reign of terror: the aftermath of 11 September 2001 and subsequent anthrax attacks has seen a shift in the focus of US research as money is ploughed into combating the threat posed by bioterrorists.



which differing interpretations of skeletal anatomy generate noisy, sometimes acrimonious debate. In October, four researchers argued that Toumai was in fact a female ancestral ape, related to modern gorillas³. The group included Martin Pickford and Brigitte Senut, both at the National Museum of Natural History in Paris, who reported in 2001 on the 6-million-year-old *Orrorin tugenensis* from Kenya⁴ — a specimen whose status as a hominid is hotly debated.

Pickford, Senut and their colleagues pointed to features such as a flat plane at the back of Toumai's skull where the neck muscles attached, taking this to indicate that the creature walked on all fours. But Brunet responded that this was a misinterpretation arising from deformation to the skull⁵, and many experts agree. With the title of finder of the oldest hominid bones at stake, the 2003 meeting of the Paleoanthropology Society in Tempe, Arizona, promises to be a lively affair.

The next argument may revolve around the right to expand expeditions into the harsh and violent land where Toumai was discovered. Previous experience elsewhere in Africa suggests the possibility of territorial scuffles between rival groups over field sites.

Brunet has said that his international team — the Franco-Chadian Palaeoanthropological Mission — is open to discussions about accepting new members. But only the toughest need apply. The region is subject to incursions of rebels and bandits from near the Libyan border, and scientific expeditions require troop escorts. But even armed guards can offer no protection against the sandstorms that regularly blast the bleak desert. ■

Rex Dalton

1. Brunet, M. *et al. Nature* **418**, 145–151 (2002).
2. Vignaud, P. *et al. Nature* **418**, 152–155 (2002).
3. Wolpoff, M. H., Senut, B., Pickford, M. & Hawks, J. *Nature* **419**, 581–582 (2002).
4. Senut, B. *et al. C. R. Acad. Sci. Paris* **332**, 137–144 (2001).
5. Brunet, M. *Nature* **419**, 582 (2002).

Biodefence

Ploughshares into swords

A year ago, Rod Welch's research was well regarded but unlikely to grab the headlines. A microbiologist at the University of Wisconsin-Madison, he worked on toxins produced by the bacterium *Escherichia coli*.

Today, Welch is in the vanguard of the 'war on terrorism'. He plans to begin new studies of the deadly toxin made by *Clostridium botulinum*, a potential bioweapons agent, and has aligned himself with a group of researchers bidding for money to build a midwestern centre for biodefence research based at the University of Chicago. He also hopes to win research grants from the pot of some \$1.75 billion that the National Institutes of Health (NIH) will award for biodefence in 2003. "There is a realization that the NIH expects people who have been trained in this area to come to bat," Welch says.

Welch is just one of thousands of US biologists who have begun recasting their work over the past year in response to the unprecedented funding opportunities for biodefence research. The bulk of the NIH's 2003 money — some \$1.36 billion — is allocated to the National Institute of Allergy and Infectious Diseases (NIAID), headed by Tony Fauci. He hopes to spend \$190 million on up to eight new regional and national biosafety labs, as well as four new regional centres of excellence.

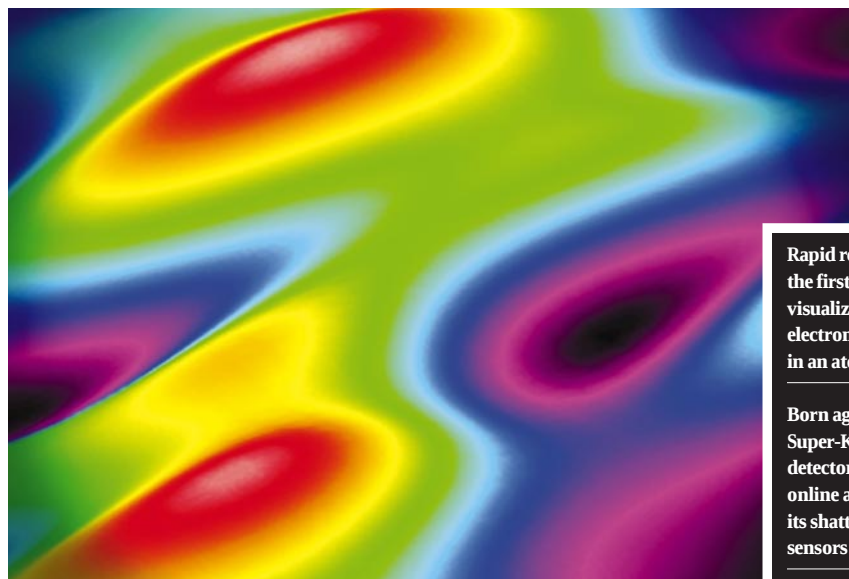
Researchers also expect the increased funding to be a boon for work on the basic biology of infectious disease. The NIAID has already called for proposals in areas such as the sequencing of microbial genomes, the development of new animal models of emerging human diseases, and the innate

immune system — the body's first line of nonspecific defence against infection. But Fauci warns that researchers must convince funders that their work will lead to new treatments, diagnostics and vaccines against bioweapons agents.

Although this initiative is already under way, there are potential roadblocks. One snag is that the NIH's paymasters in Congress have so far been unable to pass a budget for the agency, which is operating at flat funding until it receives its new appropriation. But lawmakers say that they are committed to providing the full increase for biodefence. A second unknown quantity is the new federal Department of Homeland Security, created in November. The department is supposed to work with the NIH to set priorities for biodefence research, but the details of this relationship remain unclear.

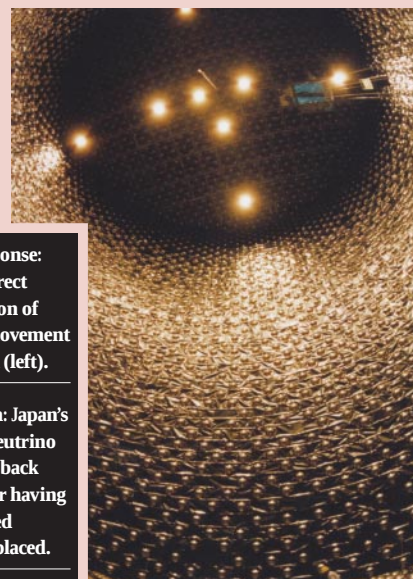
If these issues can be resolved, scientists say that new work in biodefence will pay off against a range of naturally emerging diseases — perhaps a more immediate threat than bioterrorism. This summer, the mosquito-borne West Nile virus spread across the central and western United States at breakneck speed, killing 215 people before the cool weather slowed its transmission. Officials at the Centers for Disease Control and Prevention in Atlanta, Georgia, say that improvements to the country's infectious-disease surveillance network made to counter the terrorist threat have already helped them respond better to West Nile cases this year than they had previously. Fauci is now looking to bench scientists similarly to prove their worth. ■

Erika Check



Rapid response:
the first direct
visualization of
electron movement
in an atom (left).

Born again: Japan's
Super-K neutrino
detector is back
online after having
its shattered
sensors replaced.



Attosecond science

The fast show

"It has been a fantastic year," beams Ferenc Krausz. His team at the Vienna Institute of Technology has used some of shortest laser pulses ever generated to track electrons moving within an atom. This motion is so quick that the vibrations of atoms — themselves among the fastest events subjected to scientific scrutiny — seem sluggish in comparison. Fundamental insights into atomic behaviour are likely to follow, as well as ideas for developing new kinds of laser.

For more than a decade, pulses of laser light just a few hundred femtoseconds (10^{-15} s) long have been used to follow the movement of atoms and molecules during chemical reactions. But tracking electrons requires even shorter pulses. An electron orbits a hydrogen atom in just 24 attoseconds (10^{-18} s), for example. Methods for producing suitably short pulses have emerged in the past couple of years^{1,2}, and 2002 saw their first real application.

To get a sufficiently short laser pulse, Krausz bombarded neon gas with pulses of visible light 7 femtoseconds long, which sparked collisions between the neon atoms' nuclei and their electrons. This activity generated bursts of X-ray radiation about 100 attoseconds long, which Krausz focused and combined to make a single 650-attosecond pulse.

Using such pulses, Krausz was able to track the movement of electrons around the nucleus of a krypton atom for the first time. A 650-attosecond pulse was used to initiate the Auger process — the rearrangement of electrons around an atom that follows the removal of an electron from an inner orbit.

By applying their pulse, the researchers knocked an inner electron out of its orbit, leaving an unstable hole. This caused an electron from an outer orbit to fall into the hole, which in turn displaced another electron — the Auger electron — from its outer orbit.

The researchers tracked the Auger electron using a femtosecond laser pulse, which allowed them to work out the time between the initial X-ray pulse and the expulsion of the Auger electron. They were able to pin the length of this process down to about 7.9 femtoseconds (ref. 3).

Other methods had already produced similar estimates by indirect measurements, but Krausz's experiment confirmed that sub-femtosecond pulses could be used to study the process by manipulating the electrons themselves. "This is the dream of attosecond science," says Paul Corkum, a physicist at the Steacie Institute for Molecular Sciences in Ottawa, Canada. He suggests that insights into atomic behaviour could also come from studying the way in which electrons are scattered in the collisions that create the short X-ray pulses⁴.

X-rays can also be generated when electrons move between different atomic orbitals. Krausz hopes that future attosecond studies will improve our sketchy knowledge of the processes involved. If we understood these, he argues, it might be possible to build a compact X-ray laser to study molecular structures in materials science and biology. ■

Jim Giles

1. Paul, P. M. *et al. Science* **292**, 1689–1692 (2001).
2. Hentschel, M. *et al. Nature* **414**, 509–513 (2001).
3. Drescher, M. *et al. Nature* **419**, 803–807 (2002).
4. Niikura, H. *et al. Nature* **417**, 917–922 (2002).

Comeback: Neutrino physics

Tanked up and ready for action

Yoji Totsuka had planned for a quiet 2002, as he neared the end of his stint as director of Japan's Kamioka Observatory. But those plans were shattered in November 2001, along with almost 7,000 photomultiplier sensors, when a shockwave triggered by one imploding sensor swept through the observatory's water-filled neutrino detector.

Instead of quietly drawing his tenure to a close, Totsuka has been trying to figure out what happened — and working overtime to repair the damage. "I'm still wondering why we could not anticipate that stupid accident," he says. "I don't have an answer yet."

The detector, known as Super-Kamiokande or Super-K, is some 240 kilometres northwest of Tokyo. Its headline achievement was providing evidence that neutrinos, a type of subatomic particle, possess mass. But the accident derailed other ongoing experiments, including a study of neutrinos beamed to Super-K from KEK, Japan's High Energy Accelerator Research Organization at Tsukuba, north of Tokyo.

A year of repairs has paid off, however. The detector has now been refitted with plastic-cased sensors, which should prevent a similar catastrophe, and the refilling of the 50-million-litre water tank will be finished before Christmas. The KEK experiment is expected to resume in mid-January. Super-K's sensitivity is half of what it was before the accident, but funding for a full recovery by 2006 is likely to be made available.

Totsuka is now director of KEK and has been succeeded at Kamioka by Yoichi Suzuki. "The most important thing," says Suzuki, "is to get some great results as a means of thanking the people who have supported us." **David Cyranoski**

Biometrics was no match for hair-raising tricks

People have been fooling the latest thing in security for a very long time.

Sir — Although biometric technology is believed to be a product of the hi-tech era, it is not, in fact, our generation's invention. People were using biometric technology long before the word 'biometric' was coined (see *Nature* **418**, 583; 2002; and **420**, 15; 2002). Not only that, but attempts to fool it were as common in ancient times as they are today.

The oldest written testimony of identity theft we can find dates back to biblical times, when Jacob fraudulently used the identity of his twin brother Esau to benefit from their father's blessing. *Genesis* describes a combination of hand scan and voice recognition that Isaac used to attempt to verify his son's identity, without knowing that the smooth-skinned Jacob had wrapped his hands in kidskin: "And Jacob

went near unto Isaac his father; and he felt him, and said, 'The voice is Jacob's voice, but the hands are the hands of Esau'. And he recognized him not, because his hands were hairy, as his brother Esau's hands." The false acceptance which resulted from this very inaccurate biometric test had historical consequences of unmatched proportions.

In Greek mythology, too, we are likely to find surprises. A primitive tactile sensor used by the one-eyed Cyclops after Odysseus and colleagues had destroyed his monocular face-recognition system — and which they evaded by hiding under his sheep — was actually the first biometric lock, operated more than two millennia before James Bond conquered the screen with his hi-tech gadgets.

Turning the dusty pages of the classics,

we discover a wide spectrum of biometric technologies, from voice recognition — triumphantly deceived by Dante's Gianni Schicchi, who impersonated a dead man to change a will in his own favour — to the unbeatable feature-matching face-recognition algorithm implemented by the fairy-tale heroine Little Red Riding Hood, who was unconvinced by a wolf disguised as her grandmother.

Scientists would be envious to find out that many state-of-the-art approaches in biometrics are merely a rediscovery: to quote *Ecclesiastes*, "nothing is new under the sun".

**Michael M. Bronstein,
Alexander M. Bronstein**

Department of Electrical Engineering, Technion - Israel Institute of Technology, Haifa 32000, Israel

Scientific misconduct: the state's role has limits

Sir — The Association of American Medical Colleges (AAMC) has long fostered research integrity, developed ethical codes for scientific societies, promulgated guidelines for inquiries into allegations of scientific misconduct, issued policies and recommendations for dealing with financial conflicts of interest in clinical research, and, most recently, collaborated with the Office of Research Integrity (ORI) in educational efforts and workshops on the responsible conduct of research. The AAMC has long recognized that misconduct breaches the social contract underpinning academic science and undermines a scientific establishment that sets the standard of international excellence. Contrary to your assertion, the AAMC strongly supports ethics training for graduate students and postdoctoral research fellows.

In his annual address to academic leaders and the public, the AAMC president, Jordan Cohen, stated "either we are trustworthy and deserve the privilege of self-regulation, or we are suspect and warrant the close scrutiny of government." We are proud of our record, which speaks for itself.

Your Opinion article (*Nature* **420**, 253; 2002) mischaracterizes the opposition expressed by the AAMC and the Federation of American Societies for Experimental Biology (FASEB) to the ORI's proposed data-collection instrument. Our objections focus on two critical elements: first, the likelihood that the survey would result in unusable, invalid and easily misinterpreted 'data' based on subjective criteria and

imprecise measures. One example of this is mentioned in your editorial — asking for instances where a researcher observed a colleague "citing an article they had not read firsthand". Our second objection is to the ORI's inexplicable defiance of the settled federal definition of scientific misconduct, which can lead only to confusion.

The crux of the issue, which you fail to comprehend, is that the federal definition of 'scientific misconduct' is a marker for the proper role of government in overseeing the conduct of federally funded academic research. On this matter, the US scientific community has consistently spoken with a single voice in arguing that this role be circumscribed and focused on transgressions that are reasonably unambiguous and are unacceptable across all scientific and scholarly disciplines.

This unanimity does not imply, however, that the boundaries of unethical scientific and professional behaviour should be so circumscribed — quite the contrary. It is not the role of the federal government to define or prescribe those ethical boundaries. Rather, it is the obligation of academia and scientific societies, and it is to stimulate and assist our member institutions to meet that obligation that the AAMC has been so actively and demonstrably engaged for so long.

Perhaps *Nature* believes that science needs a federally sanctioned 'high church' as the final arbiter of scientific morals, ethics and integrity. The AAMC, and, we believe, all of US science, profoundly disagrees.

David Korn

Senior Vice President, AAMC, 2450 North Street, Northwest Washington, DC 20037, USA

Scientific misconduct: ORI survey is flawed

Sir — The Federation of American Societies for Experimental Biology (FASEB) abhors misconduct in research and has repeatedly emphasized the need for clear, unambiguous and consistent definitions of misconduct. The accusations you make in Opinion (*Nature* **420**, 253; 2002) misrepresent our criticism of the US Office of Research Integrity's (ORI's) flawed survey questionnaire.

We do not object to data collection on misconduct. Institutions currently provide this information to the ORI on an annual basis. Our opinion is that the proposed ORI survey has serious deficiencies and will not produce useful data.

The ORI itself stated that previous attempts to measure misconduct were unsuccessful because they strayed from the federal misconduct definition. The issues of fabrication, falsification and plagiarism are too important to be confused with other questions, many of which involve legitimate differences of opinion. The survey's vague questions, such as asking respondents how many times they have observed colleagues "failing to cite references that contradict their current research" or "refusing to give peers reasonable access to unique research materials", will give a misleading impression of how research is done. Although it is easy to circle a number, there may be wide variation in the ethical status of the examples being reported by individuals. Simple summaries of complex issues will lump legitimate actions together

with malevolent cases.

Your editorial fails to inform readers that the ORI is not merely proposing to measure “other” misbehaviour, but also “perceived” misbehaviour, which is key to our objection. The survey will not generate a measure of misconduct, but a recording of hearsay or innuendo. For example, one question asks whether the respondent knows of colleagues “citing an article they had not read firsthand”. This Orwellian approach, which encourages scientists to spy on each other’s reading habits, will not lead to clarification of the ethical status of biomedical research.

You imply that there was a conspiracy behind the creation of the current definition of research misconduct. This is incorrect. The record clearly shows that there was an extended debate and many opportunities for public comment. We have supported efforts to improve education in research ethics, as I stated publicly in remarks at the 10 October Institute of Medicine town meeting. FASEB’s August 2000 letter commenting on the draft PHS Policy for Instruction in the Responsible Conduct of Research states: “Students and trainees must have instruction in the responsible conduct of research. But the extension of this requirement to ‘all staff’ including subcontractors and consultants will result in an enormous involvement of time and resources.” Our policy statements show our consistent commitment to the responsible conduct of research. There is no basis for implying that our position condones, supports or protects unethical behaviour.

Steven L. Teitelbaum

President, FASEB, Dept of Pathology and Immunology, Washington University School of Medicine, MS 90-31-649, 216 Kingshighway, St Louis, Missouri 63110, USA

Sounding the alarm on underwater noise

Sir — Your News feature on fisheries management (*Nature* **419**, 662; 2002) raises the important point of considering each fishery in the context of its larger ecological context, although you identify overfishing as the root cause of fisheries depletion. This perspective puts the onus of fisheries health exclusively on the fishing industry. Agricultural runoff, the pollution and disappearance of estuarine and wetland nurseries and, increasingly, anthropogenic ‘noise’ are all compromising ocean health — to such a degree that fisheries population crashes might occur even without commercial overharvesting.

Only recently has anthropogenic noise been acknowledged as a threat to marine

ecology, as evidenced by whale and dolphin strandings caused by military sonar (*Nature* **415**, 106; 2002). Noises caused by shipping, underwater seismic exploration, sonar, underwater telemetry and military exercises have all increased dramatically over the past decade. Many fish rely on acoustical perception to hunt, school, evade predators and find mates. Cluttering their acoustical niches with noise affects their survival prospects.

Michael Stocker

Seaflo.org, PO Box 559, Lagunitas, California 94938, USA

Why astronomy is the star of the news show

Sir — Space science news is good news, according to the Pew Research Center for the People and the Press, which tracks the “most closely followed” news stories in the United States (<http://people-press.org/reports>). From 1986 to 1999, the Pew study found 689 such items, 39 of them related to science, medicine and the weather. To a striking degree, these 39 stories had disturbing news to tell — earthquakes or other natural calamities, nuclear power, AIDS or controversies over cloning. Virtually every ‘good news’ science story was about space science, for example reports of the Hubble Space Telescope (1990) and Mars Pathfinder (1997) missions. The only other scientific subject reported in a positive light was Viagra.

Perhaps this ‘good news’ feature of astronomy and space science helps to explain their broad popularity. A recent NSF survey reports that 74% of adults in 1999 were interested in space exploration, and many people (57% in 1999) agreed that it is worth its costs (see “Public Attitudes Towards Space Science”, by H.A.S., in *Space Science Reviews* Vol. 102, 3–4, Kluwer, 2002.)

Politicians mindful of public morale might want to note this phenomenon, as might budget-conscious managers who seem to suspect that astronomy and space science do not give adequate value-for-dollar (and/or are badly managed). The current US administration, for example, and its Office of Management and Budget (OMB), in an effort to remedy these perceived faults, are “measuring the performance” of government-supported research so that “spending on fundamental research will be judged by formal ‘performance criteria’” (see *Nature* **415**, 466–467; 2002). In the spirit of “improving management”, the OMB first proposed to move the National Science Foundation’s astronomy funding to other agencies — with little clear reason, support or success

— and subsequently proposed the opposite: moving some astronomy (and other research) funding into the foundation (see *Nature* **414**, 680; 2002).

How does one calculate value-for-dollar in a discipline such as astronomy, whose primary product is knowledge? As succinctly (but unhelpfully) put by one astronomer, Nicholas Copernicus, the aim of a scientist is “to seek the truth in all things”. But without a sense of the worth of research accomplishments, how can those formal performance criteria for measuring and managing be identified, and then seriously and sensibly applied? Surely one lesson of the collapse of the telecommunications giant Enron is that ‘measuring’ and ‘managing’ are techniques that are themselves subject to error, incompetence, inefficiency and worse.

This is where the good-news aspect of astronomy comes in. Gerald Holton and Gerhart Sonnert (*Issues in Science and Technology*, 61–65; Fall 1999) have proposed a model in which basic research falls into three categories whose respective goals are knowledge, applied knowledge and (after Thomas Jefferson) knowledge with the realization that something practical might ensue. John F. Kennedy expressed a fourth goal: the public spirit. As he put it when justifying the Apollo programme: “We choose to go to the Moon in this decade, and do the other things, not because they are easy, but because they are hard... No single space project in this period will be more impressive to mankind.” In the words of the current NASA Administrator, Sean O’Keefe, NASA’s “mandate is to pioneer the future... NASA’s work inspires Americans and unites people.”

Public spirit and public interest, though intangible, are reflected in the media. New planets, new insights into the creation of the Universe, and other cosmic discoveries regularly merit upbeat front-page coverage because the pictures are inspirational, the stories are relatively easy to understand, the adventures are exciting, the discoveries are often meaningful, and the successes make us happy to be alive. Other science research disciplines (and their public outreach programmes) might also benefit from performance criteria that include a media-based appraisal of public attitudes. At least for astronomy the news is good — accountants, please add 10 points.

Howard A. Smith

Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA

correspondence

contributions to correspondence may be submitted to corres@nature.com. They should be no longer than 500 words, and ideally shorter.

Baby love

Cruel experiments on monkeys showed that infants need affection.

Love at Goon Park: Harry Harlow and the Science of Affection
by Deborah Blum
Perseus/Wiley: 2002/2003. 336 pp. \$26/£17.99
Alison Jolly

"You cried and cried and cried, and I wanted so much to comfort you, and I knew that I just mustn't pick you up and hold you. That's what they told me, so I believed them." That was a mother in my own family, apologizing decades later to her adult daughter. The mother still suffered because she had not dared to comfort her baby.

In *Love at Goon Park*, Deborah Blum tells the extraordinary story of Harry Harlow. It is a nuanced and brilliant evocation of a major figure in psychology who dared to challenge the orthodoxy of his time. Harlow said that babies need love, that they are born needing love, and that without contact, comfort and social responsiveness, primates, including humans, grow up as incomplete beings — if they live to grow up at all. But memories are short. Those who remember Harlow now mostly have memories of him as a man who tortured baby monkeys by removing them from their mothers and giving them surrogate wire or cloth dummies with bicycle-reflector eyes. Yes, Harlow's experiments were cruel. Yes, he knew they were cruel; he shoved that in people's faces. But he did them to show the appalling cruelty of the prevailing dogma on how to treat human children.

Blum reminds us of that dogma. John B. Watson, the founder of behaviourism, proclaimed in *Psychological Care of Infant and Child*, his best-selling childcare book of 1928: "You will soon be ashamed of the mawkish, sentimental way you have been handling your child." Behaviourists thought that we are born blank slates, formed only by conditioned reward and punishment. Milk was the reward; the baby simply learned to associate his mother with milk. Over-mothering was thought to produce over-dependent offspring. With the dawning of an understanding of hygiene, human contact was seen as passing dangerous germs. In families there was usually some compensation, but in hospitals and orphanages, babies could be kept in solitary confinement, away from care-takers and each other. Mysteriously, the infants seemed to lose interest in life, and many died. The remedy prescribed was ever more sterile isolation.

In England, the revolt against this view was led in the 1950s by John Bowlby and James Robertson. Many psychologists, though, saw the campaigners as soft, sentimental and unscientific. It was Harlow's hard science that broke the barriers of doubt. He showed conclusively that baby rhesus monkeys could not be conditioned to love a wire mother, even if she was the one equipped with a milk bottle — they clung to warm cloth instead. They would work to open a window just for a glimpse of their cloth 'mother'. They ran to the cloth mother if they were terrified by something new, such as a wind-up toy bear banging on a drum, and were comforted by 'her' presence. Monkeys raised in isolation, without contact, grew up insane, unable to deal with other monkeys. Some were unable to mate unless fastened in what Harlow, with characteristic bluntness, called a 'rape rack', and then they were likely to abuse or murder the resulting child. Of course it is obvious now. It just wasn't obvious then.

Who should read this book? Anyone working with small children, and many who are raising a small child — and anyone interested in authority in science. They should do so not just because it is beautifully and intelligently written, with a journalist's verve and a professional's attention to source and detail, but because of the questions it raises. How could the people who preceded Harlow have been so wrong? How could the antidote be a man who drove monkeys insane? Science is often blamed for its service to evil societies, Nazi eugenics or the hydrogen bombs of the Cold War, but this is a story of science unconstrained, creating its own

perversions. As Harlow put it in 1953: "It is my belief that if we face our problems honestly and without regard to, or fear of, difficulty, the theoretical psychology of the future will catch up with, and even surpass, common sense." Let's hope he was right. ■

Alison Jolly is in the School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK.

Immunology tomorrow

Immunology and Evolution of Infectious Disease
by Steven A. Frank
Princeton University Press: 2002. 352 pp.
\$75, £52 (hbk); \$24.95, £17.95 (pbk)
Franziska Michor & Martin A. Nowak

It was Charles Darwin who observed: "It is not the strongest of the species that survive, nor the most intelligent, but the one most responsive to change." In the ever-changing world of the immune system, neither size nor complexity of a parasite counts, but adaptability.

In *Immunology and Evolution of Infectious Disease*, Frank's multidisciplinary approach to infection dynamics brings together population ecology, immunology and evolution. He links evolutionary change of parasites to molecular details of immune recognition and to genetic variation of the host population. Newcomers are offered a comprehensive introduction to basic questions of immunology, as well as a synthesis that cuts across large areas of biology. Specialists will find detailed discussions of specific infectious agents from a highly original, evolutionary perspective, and inspiration for future research.

The structure of the book shows how different subjects must be integrated for comprehensive understanding of parasite variation. First, it gives a general introduction to vertebrate immunity and antigenic variation. There follows a discussion of molecular processes of immune recognition and parasite escape. Frank then focuses on the dynamics of a single infection and examines genetic variation within populations of hosts and parasites. Finally, he discusses methods to study the evolutionary dynamics of antigenic variation. As the book progresses, the complex issues of evolutionary change are related back to the structural and biochemical properties of molecular recognition introduced earlier in the book.

Antigenic variation benefits parasites by



Creature comforts: baby monkeys held on to the cloth mother (left) whenever they could.



Egon Schiele died from Spanish 'flu in 1918, days after painting *Die Familie*, which includes a self-portrait. A shift in antigen enabled the influenza A virus to evade the immune system, killing 20 million people.

allowing a longer period of infection within an individual host and by facilitating re-infection of hosts with immune memory. It also increases the abundance of parasites within an infected individual and thereby enhances infectivity. The human immunodeficiency virus (HIV) is discussed as a primary example of a parasite that evolves within a single host. Antigenic variation of HIV leads to escape from cytotoxic T cells and neutralizing antibodies. It augments viral load and thus accelerates the destruction of CD4 cells. There is a highly dynamic balance of power between HIV and the immune system, which is slowly shifted as a consequence of virus evolution to allow the virus to escape from immune recognition and to reproduce more efficiently in a broader range of different cells. Thus, virus evolution within individual hosts may well be the cause of disease progression, as was first proposed some ten years ago.

Parasites in general have found an amazing diversity of processes by which they can escape immune recognition. Mechanisms of generating antigenic variation include baseline mutation during replication, hypermutation of antigenic loci, and switching between archival variants. The analysis of infection dynamics within individual hosts includes hepatitis C virus (HCV), human T-cell leukemia virus (HTLV), and the malaria parasite *Plasmodium falciparum*.

Frank examines variability in parasites and hosts across entire populations, discussing genetic differences among hosts in their immune responses and immune memory profiles. The overview of the methods for studying evolutionary change is exceptional. Immunological and phylogenetic classification of virus variants provides evidence for selection in genetic data of influenza A virus and HIV. Experimental evolution of foot-and-mouth disease virus is discussed, as well as selection for escape from cytotoxic T cells.

This book considers the natural history of antigenic variation at the molecular, population and evolutionary level. Through its novel way of integrating all factors that affect antigenic variation in pathogen biology, it provides new insights into the evolution of infectious disease. It is a rich source of ideas for scientists working in immunology and molecular biology as well as evolution. Frank defines the key problems for the future study of parasite variation and escape from host recognition. For example, he suggests studying experimentally the relations between infection length, parasite abundance and transmission success to learn how selection shapes antigenic variation within hosts.

To determine whether antigenic selection shapes phylogeny, mathematical models can clarify the relations between antigenic and

phylogenetic classifications. Comparison of parasites in different hosts or geographical locations can reveal the effects of MHC diversity on antigenic variation.

The book brings home the point that the immunology of tomorrow will greatly benefit from making contact with structural, genomic, evolutionary and mathematical techniques. Academic institutions and granting agencies worldwide have started to predict such a multidisciplinary future for the whole of biology. Read this book and 90% of your multidisciplinary grant on any topic of infectious disease is already written. ■

Franziska Michor is in the Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

Martin Nowak is at the Institute for Advanced Study, Princeton, New Jersey 08540, USA.

A drink to your health!

Wine: A Scientific Exploration
edited by Merton Sandler & Roger Pinder
Taylor & Francis; 2002. 336 pp. £65, \$65
John Mowbray

In a mildly puritanical culture, where indulgence in alcohol causes twinges of guilt, there is a particular delight in finding that a much-used recreational drug may confer health benefits. The critical evaluation of the beneficial and possible deleterious effects of wine and alcohol consumption on humans is one theme of this multi-authored compilation on wine.

The second theme is the scientific basis and evolution of viticulture and wine production. The tale of the infection of European vines by the American phylloxera aphid and the reluctance of the French central authorities to accept the findings of compatriots from the Midi region, as well as the continuing problems with this infestation, make fascinating and instructive reading.

The currently controversial role of biotechnology in wine production and the possible benefits and limitations of genetic engineering are discussed clearly, as is the revolution in identifying the parentage and identification of vine species. The scattered contributions linking the properties of wine to its medicinal effects are covered by chapters on the production and physiological roles of polyphenols and flavonoids in the grape, and by an illuminating and succinct history of wine as a medicine.

Regarding the medical benefits of imbibing wine, there is epidemiological evidence that moderate alcohol intake protects against coronary heart disease, and possibly



against ischaemic stroke and age-related retinal degeneration. Heavy drinking and binge drinking reverses these beneficial effects, although an Italian study of heavy drinkers of wine does not support these findings. There is convincing evidence that wine has better antibacterial effects on gut bacteria and wound infection than ethanol at the same concentration. However, the stilbene derivatives, which have myriad helpful inhibitory effects *in vitro*, are conjugated as glucuronides or sulphates by the human intestinal mucosal cells and are rapidly excreted in urine, so they do not reach effective concentrations in whole animals; perhaps most of these *in vitro* studies can be consigned to “the dust heap of science”.

In places one can discern a logical progression of these two themes in succeeding chapters, but in general they are scattered throughout the book with few internal cross references. The editors say (modestly!) that the “authoritative views presented” ought to be “required reading for medical practitioners... as well as for botanists, oenologists and viticulturists, historians of science and simple wine buffs.” The contributions are indeed serious reviews of the current literature in the areas they survey, but specialist researchers in both cardiovascular medicine and viticulture will probably already have access to most of the findings elsewhere.

So will the book interest a more general reader? Probably not, given its length and price. But, given some serious professional editing, this compendium could have presented these two themes in a less rigorous but more digestible form and become a popular addition to the bookshelves of many “simple wine buffs” who have a scientific education. ■

John Mowbray is in the Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK.

London calling

A Thread Across the Ocean

by John Steele Gordon

Walker & Company/Simon & Schuster, 2002.
240 pp. \$26, £17.99

Paul M. Grant

It is quite likely that this decade will see the fulfilment of the wired and wireless global village over much of the world, each inhabitant wielding a palm-sized personal organizer with the combined power of a laptop and a mobile phone. The spectacle of teenagers

everywhere in constant communication, and our fridge texting us when it finds we've put the beef in the veggie bin, may be part of everyday life.

Such a world was virtually unthinkable barely a century and a half ago — except to a few men of great vision and energy. By the mid-nineteenth century, both America and Europe were rapidly expanding their internal communications by telegraph. But signals between the New and Old Worlds were still transmitted by ship at the speed of men, wind and steam, and an exchange of messages took at least six weeks.

In 1856, Cyrus Field, a New York merchant, began to assemble a collection of financiers, university physicists and practical engineers from America and Great Britain. Over the next ten years they would lay the first transatlantic telegraph cable between North America and the British Isles, an enterprise that ushered in the era of inter-continental communication that we take for granted today.

The chronicle of the failures, frustrations and eventual success experienced by these visionaries is brilliantly recounted by John Steele Gordon in *A Thread Across the Ocean*. A recognized historian of American business and commerce, he not only relates the engineering challenges that had to be surmounted, but delves deeply into the personalities of those involved, the geopolitical barriers placed in their way, and the desperate search for funds that followed failure after failure. This book is an extraordinary treasure for physicists and engineers, and scientists generally, who value the impact of their discipline on human affairs and respect the enormous effort demanded from them in return.

Relations between the United States and Britain were by no means cordial in the 1850s. In the American government of the

time, many remembered vividly the disastrous outcome of the war of 1812. Moreover, American claims to parts of western Canada had been thwarted by the British as recently as 1848, under threat of military reprisal.

Despite this, common ties of language, democratic culture and economic interest were enough to circumvent diplomatic and political difficulties and the first project was enthusiastically underwritten by financiers and governments on both sides of the Atlantic. Indeed, the first four attempts all involved both the US Navy and the Royal Navy.

Gordon's book is replete with maps, ocean elevations and technical drawings, and he takes us through some of the solutions to the myriad technical problems encountered, such as the brilliant application of the ballistic galvanometer by the young physicist William Thomson (better known later as Lord Kelvin) to the detection of the weak pulses of current travelling along the cable. Thomson's contributions throughout the project are an excellent illustration that “real physicists engage real problems” as well as the “purer” academic pursuits favoured by many of the present generation.

Perhaps the most important lesson to be learned from this venture is perseverance in the face of recurring failure. Imagine the embarrassment and difficulties faced by its backers when the first successful operation (on the third try) failed after barely three weeks of communication, a period that witnessed celebrations in London and New York, and the exchange of greetings between Queen Victoria and President Buchanan. But Field's raw charisma and the expectation of massive monetary returns from a properly functioning transatlantic cable ensured further attempts.

From the outset it was recognized that the



In a cartoon from 1858, Neptune is shocked by the telegraph cable laid under the Atlantic Ocean.

coordination of American and European equity markets alone would generate sufficient traffic and tariffs to quickly write off its capital investment. Moreover, its critical importance as an instrument of international policy was made apparent during this brief initial period of operation — the British government had been able to send a message to Canada that the Indian Mutiny had been put down and so the dispatch of British troops from Canada to India was no longer necessary, resulting in a significant saving in transport costs. The combination of visionary energy, national interest, an emerging technology and expectation of eventual and substantial returns to investors was then, and remains today, the greatest driving force for innovation.

And then there was luck, in this case the timely availability of the gargantuan steamer *Great Eastern*, designed and built by the

British engineer Isambard Kingdom Brunel. The *Great Eastern* was designed to be so large — five times the size of any existing vessel — that it could carry enough coal to circumnavigate the globe without stopping. By the time her keel was laid in 1857, it had already become apparent that the principal problem confronting attempts to lay the cable was its immense length and weight, and the stress that would be placed on vessels carrying it on the tempestuous North Atlantic. This prediction became stark reality, as the first three tries failed. After a chance meeting with Field, Brunel brought him to view the *Great Eastern* under construction, proclaiming, “Here is the ship to lay your cable, Mr Field.”

The *Great Eastern* never fulfilled its original intention to transform transoceanic shipping, and hence was available to Field’s company nine years later, in 1866, for the

fifth, and successful, effort to lay the cable. By this time, the cable used was three times the weight of the initial 1857 line. Without the *Great Eastern*, there would have been no transatlantic cable for many more years.

One could consume Gordon’s riveting tale in one go during a transatlantic flight between the United States and Britain, over the path of Field’s great endeavour some 13,000 metres below. But I suggest that readers take time to put the text aside now and again to contemplate the enormous consequences of this initial bonding of two nations in the nineteenth century — a bond that would strengthen over time and, as Bismarck foresaw, eventually determine the geopolitical destiny of the twentieth century. ■

Paul M. Grant is an IBM research staff member emeritus, and is at the Electrical Power Research Institute, PO Box 10412, 3412 Hillview Avenue, Palo Alto, California 94303, USA.

Science in culture

Naturally natural

Albrecht Dürer’s studies of animals have a life of their own.

Martin Kemp

No artist ever exercised a bigger impact on science than the German painter, print-maker, designer and author Albrecht Dürer. Indeed, it is fair to say that he had more effect on the course of the visually based sciences than any ‘scientist’ of his own day.

His writings on human proportion and geometry remained points of reference for centuries, above all his German treatise on “Instructions on Measurement with Compass and Ruler”, which, in its Latin translation as *Geometria* in 1527, was widely influential on three-dimensional geometry and those sciences of nature that sought geometrical foundations.

Less obvious but no less significant was the legacy he left in the natural sciences, in which he may be regarded as the true father of naturalistic illustration. His paternity is both direct, through his training of artists who were to illustrate key Renaissance texts, and general, through the example he set in his virtuoso renderings of nature in woodcuts, engravings and watercolours. Hans Baldung Grien, whose woodcuts of a brain dissection in 1541 set standards across Europe, had worked in Dürer’s workshop, and Hans Weiditz, illustrator of Otto Brunfel’s innovative *Herborum vivae icones* (1530), may be regarded as a direct follower of the Nuremberg master.

What Dürer pioneered was a graphic technique of such startling naturalism that he appeared to be able to create a ‘portrait’ of anything, whether it was a person, a hare or a mixed patch of grasses and flowers. Even a demon in Dürer’s hands looked as if it were portrayed ‘from life’.

Two astonishing studies of the *Muzzle of a Bull*, in the magnificent exhibition now on at the



The sixteenth-century painting *Muzzle of a Bull* by Albrecht Dürer has an uncannily lifelike quality.

British Museum, show his skill at its highest level. Undertaken in the early years of the sixteenth century, they may have been triggered by the phlegmatic beast in the background of his 1504 engraving of Adam and Eve, but they far transcend any preparatory function. Using a mixture of opaque body colour, inks and watercolour washes, applied with brushes of incredible fineness, he not only captures tiny morphological details — such as the shiny but cracked skin on the beast’s snout — but also infuses the specimen with an uncanny sense of life. Few artists have been able to endow naturalistic renderings with such a sense of quivering vitality.

The potential for the illustration of those sciences that were newly trumpeting their empirical foundations became clear to pioneers of the great Renaissance picture books. Two signal

works on plants and animals, Leonhard Fuchs’ *De Historia stirpium* (1542) and Konrad Gesner’s *Historiae animalium* (1551–87), depend on the techniques pioneered by Dürer to convince the spectator that we are in effect looking at the ‘real thing’, courtesy of the eye of the artist. So effective became the image that it stood in for direct experience of the specimen, which no longer need be consulted. This of course had its dangers, as Andreas Vesalius recognized. These are exemplified by Gesner’s rhinoceros, almost inevitably based on Dürer’s famed image — Dürer himself had never seen a living rhino — which came to be perceived over the centuries to be more like a rhino than the real thing.

Another question raised by Dürer’s technique centres on the much debated issue of whether the illustrator should show one particular specimen, warts and all, or a synthesis of many specimens in such a way as to represent the archetypal form. The heightened involvement that arises when we sense that we are looking at the individual subject remained a compelling device for empirically minded scientists well into the nineteenth century. After this the role of the renderer of nature was gradually, if not wholly, usurped by photography.

Not the least of the virtues of portraying things with the intensity demanded by Dürer is that it disciplines our looking. We ‘learn how to see’ through the act of drawing, as art critics such as John Ruskin maintained. This is one of the reasons why scientific drawing is not obsolete, even in this technological age. ■

Martin Kemp is in the Department of the History of Art, University of Oxford, Oxford OX1 2BE, UK.

The exhibition ‘Albrecht Dürer and his Legacy: the Graphic Work of a Renaissance Artist’ is at the British Museum in London until 23 March 2003.

Evolution and the egg

Lewis Wolpert and Eörs Szathmáry

Evolution has had at least 20 separate attempts at forming multicellular organisms. Plants, animals and fungi are the most spectacular and complex achievements of this kind. There are essentially two ways to make a simple multicellular entity out of single cells: either a single cell divides and its offspring stick together, or several solitary cells aggregate to form a colony. Division and adhesion is characteristic of multicellular forms of aquatic origin, whereas aggregation is typical in terrestrially derived colonies.

What is multicellularity? An overall coordination of at least some key function is a necessary and sufficient condition for a colony of cells to qualify as multicellular. According to this view, most bacterial colonies are not multicellular, as they lack this overall coordination, despite occasionally undergoing self-organized, patterned growth. It is the presence of developmentally differentiated cell types in a colony that makes it truly multicellular.

It is advantageous for the unit of reproduction (the propagule) to be as small as possible (that is, a single cell), as the uniformity thus created will reduce the likelihood of conflict between cells. Mutation — the engine of evolution — will upset this uniformity, and selection against mutation may favour propagules of different sizes. Mutants that affect the organism but benefit the cell (such as those that lead to cancer) cannot be effectively selected out of large propagules, so their occurrence would favour a single-celled propagule. By contrast, uniformly deleterious mutants that affect the survival of both cell and organism can be successfully selected out of a multicellular organism, so their occurrence would favour propagules that are larger than a single cell.

Yet there may be another factor at work. Development from a single cell — an egg — may have been essential for multicellular organisms to have evolved, as they have, into such an enormous variety of different forms. Could an asexual form of reproduction, involving budding (somatic embryogenesis) as in hydra, have given rise to so many complex new forms? This is not a question of competition between cells in the multicellular organism, but an issue that relates to developmental processes that generate the form of the organism. And the answer is no.

The first step in the development of a complex organism is the establishment of a pattern of cells with different states that can differentiate along different pathways. One mechanism for pattern formation is based on positional information: cells acquire a positional identity that is then converted into one of a variety of cellular behaviours, such as differentiating into specific cell types or undergoing a change in shape and so exerting the forces required for the formation of different structures. This and other patterning processes require signalling between and within cells, leading ultimately to gene activation or inactivation. Such a process can lead to reliable patterns of cell activities only if all the cells have the same set of genes and obey the same rules. During evolution, it is the change in genes that leads to the creation of new patterns of development.

In a hydra-like organism that only reproduces by asexual budding, it is impossible to evolve significant changes. There is no way that the genes in the huge number of cells involved in budding can change at the same time, and mutations in individual cells mean that they no longer share the behavioural rules of the majority. It is only through a coherent developmental programme, with all cells possessing the same genes, that organisms can evolve, and this requires an egg.

Multicellularity

Most multicellular organisms pass through a single-cell stage from which they then develop. This feature may render them more evolvable.

There are multicellular organisms, such as the cellular slime moulds, that develop by aggregation and not from an egg, but their patterns of cell behaviour have remained very simple for hundreds of millions of years. The evolution of more complex organisms increases the pressure to use an egg as a propagule.

An obvious but erroneous objection to our idea is that we are confusing the intrinsic disadvantage of asexuality, as discussed by population geneticists, with that of having a large propagule. Try to imagine a sexual life cycle that incorporates a large propagule. Each cell in the propagule would have to be fertilized by an individual sperm cell — otherwise the propagule would still be one that develops from a single fertilized egg. These multiple fertilizations would generate within-organism variation that would immediately generate a huge conflict. But this returns us to the question of whether it would be possible to evolve complex development with somatic embryogenesis that happens to be asexual.

We consider it practically impossible to have many asexual, differentiated cell lineages mutating in all sorts of directions in genetic space and yet retain the ability to evolve into viable new forms. This may not be completely impossible but, taking the broad view in evolutionary terms, organisms that develop from an egg would displace those that do not. Sexual reproduction also has long-term advantages, and follows easily and naturally once that life cycle of an organism is committed to egg production. ■

Lewis Wolpert is in the Department of Anatomy and Developmental Biology, University College London, London WC1E 6BT, UK.

Eörs Szathmáry is in the Department of Plant Taxonomy and Ecology, Eötvös University, and at the Collegium Budapest (Institute for Advanced Study), 2 Szentháromság, H-1014 Budapest, Hungary.

FURTHER READING

Bonner, J. T. *First Signals: The Evolution of Multicellular Development* (Princeton Univ. Press, Princeton, 2002).
Kaiser, D. *Annu. Rev. Genet.* **35**, 103–123 (2001).
Hammerstein, P. (ed.) *Genetic and Cultural Evolution of Cooperation*. Dahlem Workshop Rep. 90. (MIT Press, Cambridge, Massachusetts, in the press).
Maynard Smith, J. & Szathmáry, E. *The Major Transitions in Evolution* (Freeman, Oxford, 1995).
Kerszberg, M. & Wolpert, L. *J. Theor. Biol.* **193**, 535–537 (1998).



Ova development: even the extinct elephant bird's egg represented a single-celled genetic bottleneck.

Background comes to the fore

Matias Zaldarriaga

The cosmic microwave background radiation is a unique source of information about the early Universe. The detection of its polarization could lead to confirmation of an inflationary phase soon after the Big Bang.

The cosmic microwave background (CMB), discovered by Penzias and Wilson¹ in 1965, is the oldest radiation that can be observed and it permeates the Universe. The CMB is striking in its uniformity, but tiny fluctuations have been found in it that have profound implications for cosmology. Now, results from the Degree Angular Scale Interferometer (DASI) experiment, presented at the International Workshop on Particle Physics and the Early Universe in September, and described on pages 763 and 772 of this issue^{2,3}, show that the CMB is partially polarized, at the level of one part in a million — a measurement that reveals more about the very early history of the Universe.

The CMB is a relic from an early epoch when the Universe was very hot and in thermal equilibrium. As the Universe expanded, it cooled down, and today the CMB has a blackbody spectrum with a temperature of 2.728 K (ref. 4). In 1992, data from the Cosmic Background Explorer (COBE) spacecraft showed that the temperature fluctuates by ten parts per million, or 30 microkelvin, from one direction to another across the sky⁵. The DASI group now reports^{2,3} that there are differences in the intensities of orthogonal linear polarizations of the CMB photons of about 4 microkelvin which vary over angular scales of a third of a degree (Fig. 1).

The most important source for these anisotropies in temperature and polarization is the variation in the density of matter across the Universe. Gravity, an attractive force, causes these density fluctuations to grow with time. In the standard cosmological model, the anisotropies in the CMB are interpreted as a snapshot of the early stages of this growth, which eventually results in the formation of galaxies and the structure of the Universe that is observed today.

The last time that the CMB photons interacted with matter was 14 billion years ago — 450,000 years after the Big Bang. Until then, the Universe had been filled with free electrons and protons, and the CMB photons scattered very efficiently with electrons. When the intensity of the radiation incident on an electron is anisotropic, the scattering process converts some of that anisotropy into polarization of the scattered photon. Hence the CMB photons became polarized. But, as the Universe expanded, the temperature of the CMB cooled to the point at which the photons were no longer energetic

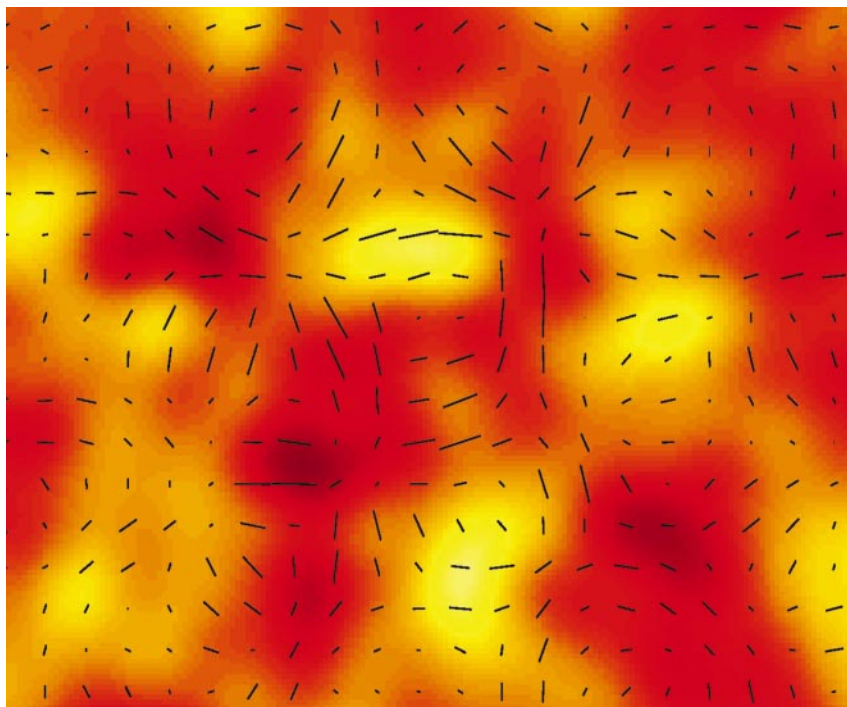


Figure 1 Polarization at the pole. The Degree Angular Scale Interferometer (DASI) in Antarctica (right) has detected^{2,3} the polarization of the cosmic microwave background (CMB). The temperature fluctuations in the CMB are shown above — yellow for hot, red for cold. Superimposed is the polarization of the CMB photons measured by DASI: the polarization at each point is represented by a black line, whose orientation and length correspond to the direction and intensity of polarization, respectively. This degree of polarization of the CMB is consistent with the cosmological model of inflation, and further measurements could reveal more about this period of rapid expansion in the early Universe.



enough to ionize hydrogen atoms and keep the Universe full of electrons and protons. This is known as the 'recombination' epoch — electrons and protons combined to form hydrogen atoms. CMB scattering with electrons stopped, so that today the CMB photons bear a unique imprint of the state of the Universe at that moment in time.

Since the discovery of temperature fluctuations by COBE, many experiments have

measured the temperature anisotropies with increasing accuracy (for a review, see ref. 6). The results have constrained the parameters in the cosmological model so much that the expected level of polarization on the angular scales probed by DASI was precisely predicted. That the DASI experiment has detected polarization at exactly this level is a resounding confirmation of the standard model.

The details of the polarization of the CMB

ERIC LEITCH

also have the potential to reveal more about cosmology and the very early history of the Universe. The density fluctuations that 'seeded' the Universe are thought to be quantum in origin, created during an early period of accelerated expansion called inflation. This inflationary period should have created a stochastic background of gravitational waves which, as they rippled through the Universe, would also have generated anisotropies and polarization in the CMB. Unfortunately, these anisotropies are too small for DASI to have observed. However, the patterns of polarization produced by density perturbations and by gravity waves are very different (the polarization vectors can only have a curl component if gravitational waves are present). Future polarization experiments could perform a unique test of inflation by identifying this gravity-wave signature. Furthermore, if inflation occurred when the Universe had cooled to an energy scale around 10^{16} GeV, the relic background of gravitational waves would actually be strong enough to be detected by CMB experiments.

The CMB polarization is also expected to contain clues about other periods in the evolution of the Universe. Almost a billion years after the recombination epoch, the first stars and quasars are thought to have started shining. The precise sequence of events in this period, usually referred to as the end of the 'dark ages', is very uncertain. The ultraviolet

radiation generated by these early sources re-ionized all the hydrogen in the Universe, providing a fresh opportunity for the CMB to scatter and become polarized — but by the time of reionization, the expansion of the Universe had diluted the density of hydrogen so much that only a few per cent of the CMB photons are thought to have been affected. Even this small fraction is enough to leave behind an observable polarization signal in the CMB. If it could be detected by future experiments, it would provide vital clues as to how and when the dark ages ended.

The detection of polarization by DASI at the level predicted by the cosmological standard model is both a remarkable technical achievement and a wonderful consistency check for the theory. In coming years, as more experiments characterize polarization on different angular scales, cosmologists hope to learn more about a variety of events in the history of the Universe. DASI has launched us on this journey.

Matias Zaldarriaga is in the Department of Physics, New York University, 4 Washington Place, New York, New York 10011, USA.

e-mail: mz31@nyu.edu

1. Penzias, A. A. & Wilson, R. W. *Astrophys. J.* **142**, 419–421 (1965).
2. Leitch, E. M. et al. *Nature* **420**, 763–771 (2002).
3. Kovac, J. M. et al. *Nature* **420**, 772–787 (2002).
4. Fixsen, D. J. et al. *Astrophys. J.* **473**, 576–587 (1996).
5. Smoot, G. F. et al. *Astrophys. J.* **396**, L1–L5 (1992).
6. Hu, W. & Dodelson, S. *Annu. Rev. Astron. Astrophys.* **40**, 171–216 (2002).

Immunology

Remembrance of things past

Michael J. Bevan

Memory T cells help us to fight off infectious microorganisms that we have encountered before. There are two models for the generation of memory cells, and new work provides support for one of them.

Vaccination is the cheapest and most effective means of protecting against certain infectious diseases. Part of this protection comes from a numbers game. In response to disease, the immune system's T lymphocytes are drafted into service: they detect and kill infected cells, and send signals that the corpses need to be removed. But killer T cells that recognize the structures (antigens) characteristic of a given infectious organism are rare in people who have not encountered that pathogen previously. Vaccination can cause the numbers of those T lymphocytes to increase many-fold, and also leads to the formation of a population of long-lived memory T cells, ensuring a rapid response to future infection. A better understanding of this process should aid vaccine development. Writing last week in *Cell*, Ahmed and colleagues¹ proposed that the production of memory killer cells is a

gradual process that may take weeks to accomplish, and requires the elimination of antigen from the body.

Killer, or cytotoxic, T lymphocytes are produced from bone-marrow-derived stem cells, which mature into precursor killer cells in the thymus and are then released into the bloodstream and lymph vessels. Each killer T cell carries the CD8 molecule on its surface, and also expresses a receptor that detects foreign antigens. These receptors vary greatly within the cell population, such that perhaps only ten cells in a million have the same antigen specificity.

When 'naïve' killer T-cell precursors encounter their cognate antigen — in the form of a pathogen or a vaccine — they multiply rapidly (Fig. 1a). As they do, they acquire the ability to kill antigen-expressing target cells and to secrete cytokine and chemokine proteins that are important in

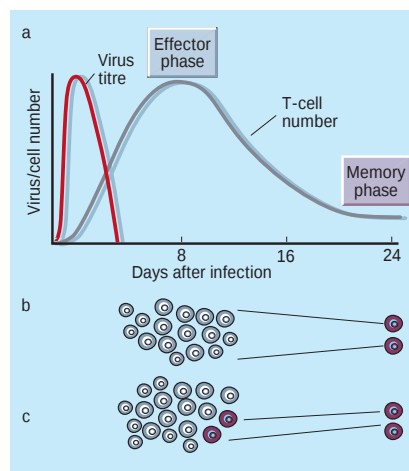


Figure 1 The T-cell response to a virus that is rapidly cleared from the body. **a**, The number of T lymphocytes that react with viral antigens increases dramatically for 8 days following infection, producing larger numbers of effector lymphocytes that kill infected cells. T-cell numbers decrease after the effector phase, leaving a stable population of memory cells. **b**, **c**, Models for how memory cells might develop. **b**, A few effector cells might be randomly selected for survival and differentiation into memory cells. **c**, Alternatively, the differentiated precursors of memory cells may already exist at the effector phase. Ahmed and colleagues' finding¹ that a few effector cells gradually convert to long-lived memory cells, even as most effector cells die, supports proposition **b**.

combating infection. They also change their migratory behaviour, allowing them to get out of the blood and lymphoid tissue and into any tissue or organ that may harbour the infection. At the peak of the response, in certain infections, activated killer T cells (effector cells) specific for the pathogen can comprise as much as 50% of the entire CD8-expressing T-cell population.

If the infection is cleared, the number of effector cells declines dramatically, and a stable population of memory T cells is left behind, often making up 1% of the CD8-expressing population. These memory cells are more common than are the naïve precursors, and can also respond more quickly following re-exposure to the antigen². The mystery is, how does the immune system 'know' to leave behind an expanded pool of long-lived memory cells? Is it that not all effector cells die when antigen is cleared (and if so, why)? Or do all effector cells in fact die off, with memory cells forming by a separate pathway that splits off early from the path leading to full-blown effectors? Ahmed and co-workers¹ aimed to find out which of these models (Fig. 1b, c) is correct.

The authors started by studying the response of mice to a short-lived (acute) infection with lymphocytic choriomeningitis virus, isolating virus-specific killer

T lymphocytes at 8 days after infection (effector cells) and at later time points as memory cells developed. They probed these cell populations by analysing gene expression and function, searching for clues to when memory cells appear.

They found that the expression profiles of effector and the later, memory cells overlap, suggesting that memory cells derive from effector cells and are not a separate lineage. The results also support the interpretation that memory cells do not exist at the peak of the effector response, but appear gradually over three weeks after the antigen has gone. The results of the functional and expression analyses — carried out at various time points between the peak of the effector stage and over 40 days after the infection was cleared — imply that the properties of effector cells fade gradually as the hallmark properties of memory cells are acquired. This slow conversion occurs even as the number of antigen-specific cells declines.

Ahmed and colleagues¹ also tested the ability of effector and later populations to handle a second encounter with the same pathogen: only the late populations did this effectively. The explanation for the superiority of late memory cells may be provided by the authors' demonstration that only these cells can undergo a second explosive population expansion in response to re-encounter with the antigen. Clearance of the antigen after the first acute infection, resulting in the cessation of signalling from the T-cell antigen receptor, apparently allows a fraction of effector cells to 'reload' to respond to re-infection.

This type of immunological memory, however, probably protects only against acute infections, such as that studied by Ahmed and colleagues, and not against chronic infections by pathogens that find ways to coexist with the immune system. In diseases such as AIDS, tuberculosis and leprosy, the differentiation of this kind of memory population may not be permitted, because antigen is not removed. Control of these infections probably depends instead on the continuous presence of effector cells³.

Recently, it has become apparent that memory T cells exist in two flavours. First, there are central memory cells that circulate through the lymphoid system just like naive T cells specific for the antigen, but in greater numbers and with a more rapid response rate. And second, there are 'effector' or 'tissue' memory T cells that can leave the blood and access all tissues where they might meet antigen (mucosal tissue, for example), before the pathogen has a chance to disseminate⁴. It remains to be seen whether these memory T cells represent distinct pathways of differentiation from effector cells, or different stages along a single pathway — although another recent report⁵ suggests the latter possibility, that effector memory T cells convert into central memory cells.

And it is still a mystery as to how a fraction of effector cells 'decides' not to die but to become memory cells. What is clear is that the more we learn about the progression of the immune response, the better able we will be to design strategies for effective vaccination. Judging by the present results¹, it is clear that 'booster shots' for vaccinations should be timed for the period when the first antigen has been cleared, and a stable population of memory cells has developed. ■

Condensed-matter physics

Dense ice in detail

Dennis D. Klug

A substance as simple as water is in fact a rich source of interesting physics. The solid phase contains several amorphous forms of ice, including one with a very dense structure, the details of which have now been revealed.

Ice is famously anomalous — unlike most solids, in its usual form it is less dense than liquid water. In fact, regular ice has so far been found to come in at least 13 different forms. Furthermore, many types of amorphous ice, without the regular molecular structure of normal ice, have been discovered. Writing in *Physical Review Letters*, Finney *et al.*¹ describe the structure of a new, very dense form of ice, whose existence has profound implications for many aspects of research into amorphous ice and other solids, and for our understanding of the phase diagram of water and ice (Fig. 1).

This 'very-high-density amorphous' (VHDA) ice, 40% denser than ordinary ice, was originally discovered by Loerting *et al.*² when they modified the conventional recipe for preparing high-density amorphous ice, commonly denoted HDA. The HDA form does not have the hexagonal crystal structure of ordinary ice, and is nearly 30% denser: each water molecule is surrounded by roughly five nearest-neighbour water molecules, whereas in hexagonal ice each molecule has exactly four nearest neighbours.

The preparation of HDA was first reported in 1984 by Mishima *et al.*³, by a method called pressure-induced amorphization: samples of crystalline hexagonal ice are pressurized at low temperatures (usually 77 K) and, at a pressure of about 1 gigapascal (GPa), a very sharp transformation occurs as the ice takes on the HDA form. The HDA ice can be brought back to ambient pressure as long as the temperature remains low. But if it is warmed to about 120 K at ambient pressure, it transforms into low-density amorphous (LDA) ice, releasing a fair amount of heat⁴. A series of amorphous-ice forms with densities between that of HDA and LDA have since been characterized⁵. When Loerting *et al.* modified the Mishima recipe and heated HDA up to about 160 K,

Michael J. Bevan is in the Department of Immunology, Howard Hughes Medical Institute, University of Washington, Box 357370, Seattle, Washington 98195-7370, USA.
e-mail: mbevan@u.washington.edu

1. Kaech, S. M., Hemby, S., Kersh, E. & Ahmed, R. *Cell* **111**, 837–851 (2002).
2. Sprent, J. & Surh, C. D. *Curr. Opin. Immunol.* **13**, 248–254 (2001).
3. Zinkernagel, R. *Curr. Opin. Immunol.* **14**, 523–536 (2002).
4. Sallusto, F., Lenig, D., Forster, R., Lipp, M. & Lanzavecchia, A. *Nature* **401**, 708–712 (1999).
5. Wherry, E. J. *et al. Nature Immunol.* (in the press).

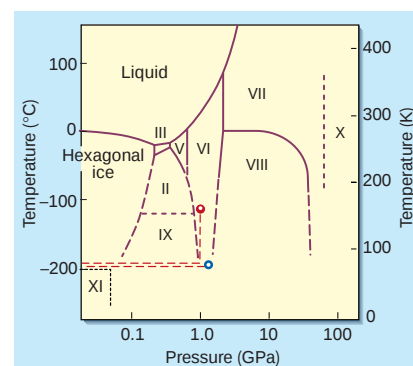


Figure 1 The phase diagram of ice. The temperature and pressure regimes associated with most of the 13 known crystalline phases are indicated here. When hexagonal ice at 77 K is subject to increasing pressure, so-called amorphous ice forms: at 1 GPa (blue circle), high-density amorphous ice forms; if the temperature is then raised, very-high-density amorphous ice forms (red circle).

maintaining the pressure at 1.15 GPa, they noticed that the ice became even denser — VHDA had formed.

Why is VHDA important? The original discoveries of HDA and LDA led to the suggestion that HDA and LDA were related to supercooled forms of liquid water. The idea was based mainly on the observation that HDA was formed at high pressure close to the extrapolation of the curve that describes how the melting point of hexagonal ice changes with pressure. The melting curve of ice has a negative slope and ice in fact melts at -20°C under a pressure of 0.2 GPa, so this was a reasonable initial suggestion. There has been speculation that, as HDA and LDA exist, there may be two forms of liquid water at low temperatures. A well-defined boundary between these phases could mean that there is a second critical point, for transitions

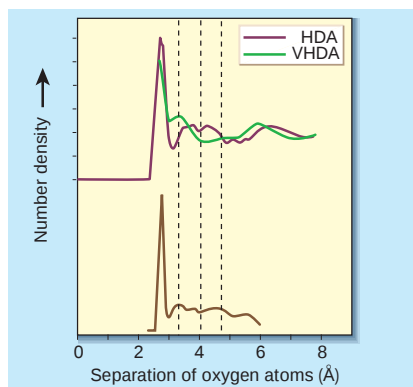


Figure 2 Molecular spacing. For an oxygen atom in a water molecule, the radial distribution of neighbouring oxygen atoms measured by Finney *et al.*¹ is shown (top) for high-density amorphous ice (HDA) and very-high-density amorphous ice (VHDA). The lower plot shows the same distribution derived from computer simulations of dense amorphous ice formed through mechanical collapse of crystalline ice¹⁰, and is a reasonable match to the VHDA plot. (Graphs redrawn from refs 1, 10.)

between the two liquid phases, in the phase diagram for water (the other critical point being at a temperature of 374 °C and a pressure of 22 MPa). From experiments and theoretical calculations, this critical point has been predicted⁶ to be at 220 K, for a pressure of 0.1 GPa. Clearly the discovery and characterization of new dense forms of amorphous ice alters the landscape for predicting and locating a liquid–liquid transformation and a new critical point, and may even call the hypothesis into question.

To understand the structure of VHDA, Finney *et al.*¹ have done a three-dimensional structural analysis using neutron diffraction and computer simulations to determine whether VHDA can be related to dense water. They found that each water molecule is surrounded by roughly six other water molecules — in hexagonal ice the number is just four. This high coordination number in VHDA perhaps results from annealing of the high-pressure amorphous structure as the temperature is raised.

There is another possibility. It seems that the transition from hexagonal ice to HDA at low temperature may in fact be more complex than a conventional thermodynamic phase transition, possibly resulting instead from a mechanical collapse of the ice lattice⁷. The discovery of VHDA ice makes this transition even more interesting, and early results⁸ of computer simulations show a strong resemblance between a collapsed and densified ice form and VHDA (Fig. 2). It might be that a transition to VHDA at 77 K is inhibited for simple kinetic reasons, as heating to 160 K is required to reach the region of phase space in which VHDA can form. It is interesting to note that the recipe for preparing VHDA

is rather specific: minor variations in the pressure and temperature conditions yield a crystalline phase that is also dense and has similar spectroscopic features^{2,9,10}.

Our understanding of the phase diagram of water will certainly undergo a critical re-examination as a result of the discovery of VHDA and the elucidation of its structure by Finney and colleagues. The process by which VHDA is formed, pressure-induced amorphization, should also be investigated further to understand exactly how this unusually dense phase of water occurs. ■

Neurobiology

How hardwired is the brain?

Ole P. Ottersen and P. Johannes Helm

In a technological breakthrough, two groups have shown that it is possible to study the turnover of spines — tiny protrusions on nerve cells — in live mice. But it's still uncertain just how dynamic the spines are.

As everyday experience tells us, our brains have a remarkable capacity to adapt to changes in the environment and to form memories. But to what extent does this malleability of the adult brain depend on physical rearrangements of the connections between nerve cells? This is one of the most challenging questions in neurobiology, and is addressed by Trachtenberg and co-workers¹ and Grutzendler and colleagues² on pages 788 and 812 of this issue. Using a combination of technologies, the authors were able to view individual neurons in the cerebral cortex of live mice.

Both groups^{1,2} focus on neuronal spines (Fig. 1). These are tiny protrusions from long, slender extensions (dendrites) of nerve cells, and they constitute the receiving parts of synapses — the contacts that mediate neuron-to-neuron signalling. Spines are believed to be the most basic functional units in the brain³, and their large number (of the order of 10⁴ per neuron) helps to explain the impressive memory-storage capacity of the cerebral cortex. Trachtenberg *et al.*¹ and Grutzendler *et al.*² have investigated whether spines are stable over time, or whether they retract and reappear — a pivotal question for our understanding of the brain's plasticity and its ability to form and store memory traces.

The analysis of structures as small as spines (which have diameters in the sub-micrometre range) is the domain of light and electron microscopy. As post-mortem tissue provides a static picture of the brain, neuroscientists have examined cell cultures or brain slices as substitutes for the intact brain. Such studies indicate that spines are dynamic structures^{4,5} and may grow in response to intense synaptic stimulation^{6,7}. But it is uncertain to what extent such *in vitro*

Dennis D. Klug is at the National Research Council of Canada, Ottawa, Ontario K1A 0R6, Canada. e-mail: dennis.klug@nrc.ca

1. Finney, J. L. *et al.* *Phys. Rev. Lett.* **89**, 205503 (2002).
2. Loerting, T. *et al.* *Phys. Chem. Chem. Phys.* **3**, 5355–5357 (2001).
3. Mishima, O. *et al.* *Nature* **310**, 393–395 (1984).
4. Handa, Y. P., Mishima, O. & Whalley, E. *J. Chem. Phys.* **84**, 2766–2770 (1986).
5. Tulk, C. A. *et al.* *Science* **297**, 1320–1323 (2002).
6. Mishima, O. & Stanley, H. E. *Nature* **396**, 329–335 (1998).
7. Tse, J. S. *et al.* *Nature* **400**, 647–649 (1999).
8. Tse, J. S. & Klein, M. L. *Phys. Rev. Lett.* **58**, 1672–1675 (1987).
9. Loerting, T. *et al.* *J. Chem. Phys.* **116**, 3171–3174 (2002).
10. Salzmann, C. *et al.* *J. Phys. Chem. B* **106**, 1–6 (2002).

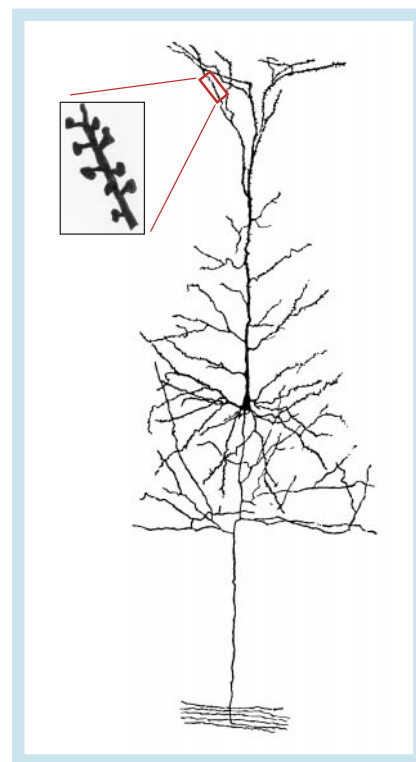


Figure 1 A century of fascination with neuronal spines. This diagram, drawn more than 100 years ago by Santiago Ramón y Cajal, shows a pyramidal cell from the mouse cerebral cortex. We have added the inset to indicate how dendrites are studded with small protrusions called spines, the subject of the studies by Trachtenberg *et al.*¹ and Grutzendler *et al.*². Like the cells analysed in these two studies, the cell drawn by Ramón y Cajal had its cell body (centre) deep in the cortex and dendrites extending towards the cortical surface (top). (Diagram modified from ref. 11.)

models are representative of the living brain.

The use of light microscopy for studying the turnover of spines in the cerebral cortex of live animals has previously seemed only a pipedream. But Trachtenberg *et al.* and Grutzendler *et al.* now show that spines of cortical neurons can be imaged over time, providing the first estimates of the rate of spine turnover *in vivo*. This technological breakthrough will pave the way for further studies of the morphology and physiology of neurons in the intact cerebral cortex.

Both groups started by 'engineering' mice to express a fluorescent protein in a small fraction of neurons in the cerebral cortex, allowing the neurons to be tracked over time. But several technical hurdles had to be overcome before the labelled neurons could be imaged. In particular, using visible light to peek into an animal's cerebral cortex through a window in its skull is fraught with difficulties, as the light is absorbed and scattered by living tissue. Infrared light, in contrast, can penetrate several hundreds of micrometres into the brain — far enough to reach the upper layers of the cortex. But the low-energy photons of infrared light do not easily excite the fluorescent proteins used to label neurons.

Enter multiphoton absorption — a phenomenon described theoretically as early as 1931 by Maria Göppert-Mayer⁸. She predicted that optical electronic transitions in atoms or molecules can be achieved even in response to low-energy (infrared) photons, provided that several photons are absorbed at the same time. In other words, if the photon flux is high enough to allow simultaneous or near-simultaneous 'hits', the energy of several photons combines to provide excitation. This prediction was borne out after the invention of the laser, and in the late 1980s Webb and colleagues⁹ showed that multiphoton absorption in fluorescent biological tracers can be achieved in the focal region of a strong microscope objective.

Trachtenberg *et al.*¹ and Grutzendler *et al.*² used the same, multiphoton-based imaging approach. However, they come to different conclusions about the stability of spines in adult animals. Trachtenberg *et al.* identified three populations of spines on superficial dendritic branches — transient, semi-stable and stable spines — with different turnover rates. Even the stable pool (representing about 60% of spines) had a limited lifetime, of around 100 days. In contrast, Grutzendler *et al.* found that about 96% of spines remained stable over a month, translating to a half-life of more than 13 months. This implies that synapses could persist throughout a mouse's lifetime.

What could explain this discrepancy? The two groups examined different cortical areas, the barrel cortex¹ and the primary visual cortex²; these might well differ in their degree of structural plasticity. And

the authors generally used adult mice of different ages: Trachtenberg and colleagues studied 5–10-week-old mice, whereas those analysed by Grutzendler *et al.* averaged 4.2 months old. But Grutzendler *et al.* also studied young mice (aged 1 month), and again found a high degree of spine stability. So the divergent results cannot be explained solely by age differences, and the reason for them remains uncertain.

Trachtenberg *et al.* also investigated several other issues concerning the role of spine dynamics in adult brain plasticity. For example, they used serial-section electron microscopy to demonstrate that new spines actually contribute to synapses; in fact, new spines that are part of morphologically mature synapses could appear from one day to the next. Trachtenberg *et al.* also find that the turnover of spines seems to be influenced by sensory experience: if the animals' whiskers are trimmed — which changes the sensory input to the barrel cortex — the pool of transient spines increases significantly. The authors also imaged dendrites and axons (another type of neuronal extension), and found that they were totally stable, in striking contrast to the situation in the developing barrel cortex¹⁰. They therefore conclude that, in the adult cortex, structural plasticity is restricted to spines.

So, how hardwired is the adult brain? The lack of consensus about the lifetime of spines^{1,2} means that we still have no clear answer. But the technological advances made

here suggest that the solution is within reach (although possible confounding effects of temperature increase and phototoxicity in the focal region must be ruled out). If it turns out that spines are highly stable, as Grutzendler *et al.*² propose, they may provide a substrate for long-term memory. But if spines are short-lived — as Trachtenberg *et al.*¹ conclude — this would challenge our understanding of long-term memory. It would also constrain the relevance of models of learning that are based on changes in existing synapses, such as long-term potentiation and depression. Whatever the case, the fact that imaging of the living brain has entered the microscopic arena represents a breakthrough that will have far-reaching implications for neurobiology. ■

Ole P. Ottersen and P. Johannes Helm are at the Centre for Molecular Biology and Neuroscience, and the Department of Anatomy, Institute of Basic Medical Sciences, University of Oslo, NO-0317 Oslo, Norway.

e-mail: o.p.ottersen@basalmed.uio.no

1. Trachtenberg, J. T. *et al.* *Nature* **420**, 788–794 (2002).
2. Grutzendler, J., Kasthuri, N. & Gan, W.-B. *Nature* **420**, 812–816 (2002).
3. Yuste, R. & Denk, W. *Nature* **375**, 682–684 (1995).
4. Harris, K. M. *Curr. Opin. Neurobiol.* **9**, 343–348 (1999).
5. Matus, A. *Science* **290**, 754–758 (2000).
6. Maletic-Savatic, M., Malinow, R. & Svoboda, K. *Science* **283**, 1923–1927 (1999).
7. Engert, F. & Bonhoeffer, T. *Nature* **399**, 66–70 (1999).
8. Göppert-Mayer, M. *Ann. Phys.* **9**, 273–294 (1931).
9. Denk, W. *et al.* *Science* **248**, 73–76 (1990).
10. Lendvai, B., Stern, E. A., Chen, B. & Svoboda, K. *Nature* **404**, 876–881 (2000).
11. Ramón y Cajal, S. *Histologie du Système Nerveux de l'Homme & des Vertébrés* (Maloiné, Paris, 1911).

Applied physics

A light-emitting sandwich filling

Tetsuo Tsutsui

Electroluminescence has taken a new turn with the combination of organic light-emitting diodes and inorganic quantum dots. The efficiency of light emission is much higher in the latest easy-to-make device.

On page 800 of this issue, Seth Coe and colleagues¹ report the fabrication of high-efficiency organic light-emitting diodes (LEDs) in which the light-emitting centres are cadmium-selenium (CdSe) nanocrystals, or quantum dots. Organic LEDs bring with them the advantages of robust fabrication technique and high performance, which, when coupled with the excellent luminescent properties of nanocrystals, offer exciting prospects for real, workable devices.

Mobile phones with small colour displays using organic LEDs are already commercially available. The images are generated through fluorescence, as electrons make transitions between orbital states of π -conjugated organic molecules (the π -bond arises from the overlap of the $2p$ orbitals of electrons in carbon atoms). As well as having high

quantum efficiency for electron-to-photon conversion, π -conjugated molecules in organic LEDs have the advantage of colour tunability, so that they can be used to build full-colour displays of red–green–blue (RGB) emitters.

But there is also a drawback: the emission spectra of π -conjugated molecules are very broad, typically spanning 50 to 100 nm (the 'full width at half-maximum', or FWHM; Fig. 1, inset). This range of molecular fluorescence is caused by the vibrational and rotational motion of atoms inside the π -conjugated molecules. So with an organic LED, it is difficult to get, for example, pure red light emitted with high quantum efficiency. Nevertheless, sharp RGB emission has been demonstrated from LEDs based on some specific materials, such as europium chelates, aggregated structures of cyanine dyes or

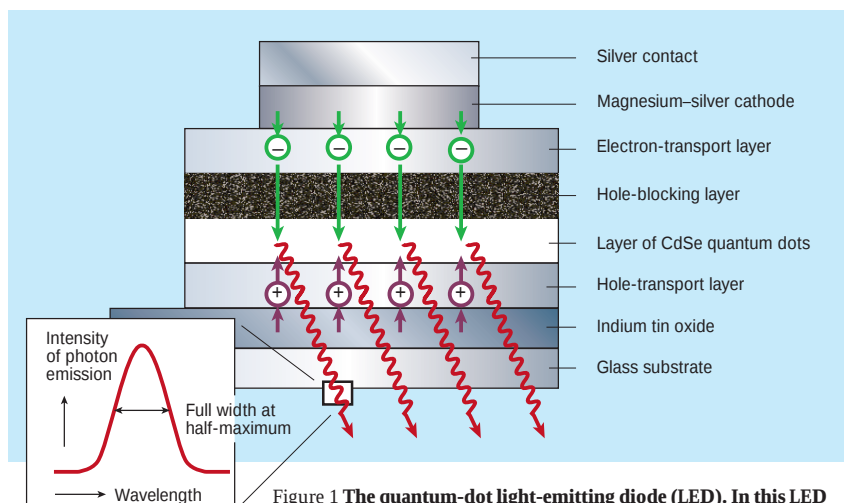


Figure 1 The quantum-dot light-emitting diode (LED). In this LED made by Coe *et al.*¹, a layer of cadmium-selenium nanocrystals, or

quantum dots, is sandwiched between layers of electron-transporting and hole-transporting organic materials. An applied electric field causes electrons and holes to move into the nanocrystal layer, where they are captured in the quantum dots and recombine, emitting photons. The spectrum of photon emission is narrow, characterized by its full width at half the maximum value.

layered inorganic-organic perovskite compounds. But these LEDs have not achieved the emission efficiency or device durability needed for practical display applications^{2,3}.

Some inorganic nanocrystals emit visible light with sharp emission spectra that are less than 30 nm FWHM. The nanocrystals are in effect quantum dots and confine charge so well within their small volume that a high quantum efficiency, exceeding 50%, is possible. It might be expected, then, that quantum dots incorporated in organic LEDs would make excellent emission centres⁴. In fact, electroluminescence has been observed by simply mixing inorganic nanocrystals with

π -conjugated polymers, but the emission efficiency was far lower than that of conventional polymer LEDs^{5,6}.

Coe *et al.*¹ have fabricated an organic LED with a single layer of CdSe quantum dots sandwiched between organic thin films (Fig. 1). Remarkably, the efficiency of their device is about 25 times higher than that achieved so far with quantum-dot LEDs. Thinking ahead to the design of device architectures, there are two noteworthy aspects here: the structure of this quantum-dot LED is already close to that for an optimum device and the process of fabricating a layer of quantum dots is simple.

For the structure of LEDs, one of the most challenging design aspects is how to bring electrons and holes together in small regions so that they recombine efficiently to emit photons without escaping or dissipating. The favoured structure is made of three layers: a thin emissive layer sandwiched between a hole-transport layer (HTL) and an electron-transport layer (ETL). If the emissive layer is thick, electrons and holes must be injected into it and transported; the emissive layer must then have the capabilities of both the ETL and the HTL, which is not ideal. If instead the emissive layer consists of a single layer of molecules, electrons and holes may be transferred directly from the surfaces of the ETL and the HTL, and high recombination efficiency is expected. At least two examples of organic LEDs with a molecular-size emissive layer have been reported: a molecular bilayer of cyanine dyes⁷, and a rubrene dye layer⁸, each sandwiched between an ETL and HTL of 50-nm thickness.

The design of Coe and colleagues' LEDs follows this lead (Fig. 1). The emissive layer in their quantum-dot LED is only a few nanometres thick, consisting of uniformly distributed single nanocrystals, each about 3 nm in diameter. The array of quantum dots was easily prepared by self-assembly in a process known as spin-casting: a solution of nanocrystals in an organic material is poured onto a substrate, which is then set spinning to spread the solution evenly. There is then a spontaneous phase segregation, as the nanocrystals pop up onto the top of the organic layer.

Coe *et al.*¹ propose that electrons and holes are captured directly at the surfaces of the CdSe nanocrystals — into discrete

Plant science

On the slide

The pitcher plant *Nepenthes alata*, which inhabits the forests of Indonesia, has one of the most dramatic organs of any carnivorous plant, as the picture here attests. In this month's *New Phytologist* (**156**, 479–489; 2002), Laurence Gaume and colleagues describe their investigations of the features that make this vessel so efficient at trapping insects.

Gaume *et al.* tested the ability of *N. alata* to capture ants and a flightless strain of fruitfly. Victims were placed on the pitcher's ridged rim. Few escaped, because any insects that stepped onto the waxy zone covering half of the

pitcher's inner surface almost inevitably fell to their doom. Ants managed to hang on to the waxy zone slightly longer than flies, and rather than fall straight into the pool of digestive fluid at the bottom of the pitcher, some landed on the layer of glandular cells that coat the bottom half of the pitcher wall. But the cells secrete a viscous substance that prevented these ants, or the few that managed to clamber out of the digestive pool, from escaping their fate as insect stew.

Electron-microscopic investigation of the surfaces provided clues to their properties. In the case of the

waxy layer, the covering is flaky and so breaks off easily — explaining why flies, which can normally cling tenaciously to a smooth surface, fell from it faster than ants. But that cannot be the whole story, as ants still had difficulty climbing surfaces from which wax had been removed with chloroform.

This study illuminates how the cunning adaptations of *N. alata*'s epidermal surfaces make the plant deadly to almost every insect that strays under its pitcher's lid. It might also inspire chemists to consider uses for materials with similarly exotic surface properties. **Christopher Surridge**



DAN SAMS/A-Z BOTANICAL

electronic levels of the quantum dots — to produce recombination luminescence. There is another explanation, that energy is transferred directly from the excited states of molecules in the surface region of the ETL or HTL. But the data collected by Coe *et al.* support the model of direct electron-hole capture and emission within the nanocrystals, and, from the perspective of high efficiency and stability of emission, this is very promising. In addition, the spread, or bandwidth, of the emission spectrum from these devices has an almost perfect gaussian profile and is determined by the uniformity of the quantum-dot size, which can be accurately controlled during fabrication. And by varying the actual size of the quantum dots, the luminescent spectrum can be tuned to particular wavelengths.

The authors also mention another advantage of this quantum-dot LED. In fluorescent materials, including the polymers used in conventional LEDs, the restrictions of statistical mechanics mean that fewer than half of the electron-hole recombinations result in light emission. This is the principal reason for the low quantum efficiency of

such devices⁹. But in inorganic nanocrystals every electron-hole recombination can produce a photon. So, in addition to the newly discovered possibilities of phosphorescent emitters¹⁰, building quantum-dot LEDs with simple inorganic nanocrystals opens a new route towards achieving a quantum efficiency of 100% at any visible wavelength. ■

Tetsuo Tsutsui is in the Department of Material Science, Faculty of Engineering Sciences, Kyushu University, Kasuga, Fukuoka 816-8580, Japan.
e-mail: tsuigz@mbox.nc.kyushu-u.ac.jp

1. Coe, S., Woo, W.-K., Bawendi, M. & Bulović, V. *Nature* **420**, 800–803 (2002).
2. Era, M., Hayashi, S., Tsutsui, T. & Saito, S. *J. Chem. Soc. Chem. Commun.* 557–558 (1985).
3. Era, M., Morimoto, S., Tsutsui, T. & Saito, S. *Appl. Phys. Lett.* **65**, 676–678 (1994).
4. Hines, M. A. & Guyot-Sionnest, P. *J. Phys. Chem.* **100**, 468–471 (1996).
5. Dabbousi, B. O., Bawendi, M. G., Onitsuka, O. & Rubner, M. F. *Appl. Phys. Lett.* **66**, 1316–1318 (1995).
6. Mattoussi, H. *et al.* *J. Appl. Phys.* **83**, 7965–7974 (1998).
7. Era, M., Adachi, C., Tsutsui, T. & Saito, S. *Chem. Phys. Lett.* **178**, 488–490 (1991).
8. Matsumura, M. & Furukawa, T. *Jpn. J. Appl. Phys.* **40**, 3211–3214 (2001).
9. Tsutsui, T. *Mater. Res. Soc. Bull.* **22**, 39–45 (1997).
10. Adachi, C. *et al.* *J. Appl. Phys.* **90**, 5048–5051 (2001).

Genomics

Return of a little squirt

Henry Gee

The draft sequence of the genome of a dim and distant relation of vertebrates will allow closer inspection of vertebrate origins. Some people have been waiting more than a hundred years for this.

The sea squirt *Ciona intestinalis* is an unassuming creature whose wider significance lies largely in its ancestry: it is a deuterostome, a member of the great group of multicellular animals that also includes the vertebrates, and so ourselves. The draft sequence of its genome, published last week in *Science*¹, promises to shed light on the evolutionary origins of vertebrates and their relationships with other creatures.

Of the animals whose genomes have been published, the roundworm *Caenorhabditis elegans* and the fruitfly *Drosophila melanogaster* are protostomes, members of the other great group of multicellular animals, and therefore only remotely related to vertebrates. Humans, mice and the puffer fish *Fugu rubripes* are all bona fide vertebrates, and so their genomes can tell us little about the sequence of events that led to the evolution of the extraordinary suite of features that make vertebrates so distinctive. These include the heart, the nervous system and lymphocyte-based adaptive immunity — features that are so characteristic, and in many ways so complex, that tracing their origins among the altogether simpler invertebrates has proved an insurmountable challenge.

What is required is the genome sequence of an animal that is a deuterostome but not a vertebrate, and which shared an ancestry with vertebrates that excludes protostomes. Enter *Ciona* (Fig. 1, overleaf), the first non-vertebrate deuterostome to have its genome published in any form. The sequence is preliminary, and mapping data are too sparse to allow the sequence to be placed on *Ciona*'s 14 chromosomes. But the researchers¹ can say that *Ciona* has just under 16,000 genes, comparable to the number for *Caenorhabditis* (19,000) and *Drosophila* (14,000) but less than the 30,000-plus seen in vertebrates. And with a genome of around 160 million base pairs — a twentieth the size of the human genome — genes in *Ciona* are as densely packed as they are in protostomes.

The main message is that the sea-squirt genome looks like a vertebrate genome, only simpler. *Ciona* has single copies of genes that are not found in protostomes, but which in vertebrates have diversified into entire families involved in cell signalling and developmental regulation: families such as the fibroblast growth factors, Smads and T-box genes. *Ciona* contains similar genes to those that, in vertebrates, direct distinctively



100 YEARS AGO

Though Jupiter has been unfavourably placed for European observers during the present year, his surface markings have been extremely interesting, of great variety and in plentiful numbers. The English climate, even at its best, can scarcely be said to suit astronomical work in an eminent degree, but its characteristics in 1902 have proved unusually bad... The most noteworthy incident in connection with recent studies of Jupiter is to be found in a very pronounced acceleration of motion in the great red spot. This first made itself evident in 1901, but it has been intensified during the past summer. For about twenty-three years, uninterruptedly, this singular marking had exhibited a constantly increasing retardation, which caused its rotation period to lengthen from about 9h. 55m. 34s. to nearly 9h. 55m. 42s. But in 1901 it declined to 9h. 55m. 41s. and during the present year the rate has been about 9h. 55m. 39½s. And this increase in velocity has been contemporary with the outbreak of a large, irregular or multiple marking of a dusky hue, in the same latitude of the planet.

From *Nature* 18 December 1902.

50 YEARS AGO

The existence of discrete sources of extra-terrestrial radio-frequency radiation is now well established and the positions of more than one hundred sources have been published. Attempts to identify these sources with any particular class of visual object have so far failed, and the origin of the radiation remains unexplained. One of the fundamental requirements in the study of these sources is a knowledge of their apparent angular size, and although attempts to make this measurement have been made by several observers, it has proved to be beyond the resolving power of their equipment. The present communication gives a preliminary account of a successful attempt to measure the angular size of the two most intense sources... The measurements are not yet adequate to define satisfactorily the shape of the sources or the distribution of intensity across their disks. Further observations are now being made... We wish to thank Dr. R. Q. Twiss for his assistance with the mathematical theory, and Prof. A. C. B. Lovell for making the necessary facilities available and for his interest in the investigation.

R. Hanbury Brown, R. C. Jennison
& M. K. Das Gupta

From *Nature* 20 December 1952.

vertebrate processes of development, such as formation of the nervous system and heart. All of this suggests that the sea squirt retains a primitively simple complement of genes that would have been found in the last common ancestor of sea squirts and vertebrates, more than 550 million years ago.

The converse is also true — *Ciona* has had plenty of time to evolve its own, non-vertebrate peculiarities. Sea squirts have some interesting physiological quirks, the most bizarre of which is the manufacture of cellulose for their distinctive outer coverings, or 'tunics'. Some of the enzymes involved in cellulose metabolism are undoubtedly home-grown, but the report¹ of a cellulose synthase gene, previously found in plants, is unique for animals and might have been an import from nitrogen-fixing bacteria.

Genomes of other non-vertebrate deuterostomes will be required to fill in the fine details of the order in which the various distinctively vertebrate developmental and physiological features were acquired. In this context, the genome of the lancelet *Branhiostoma* — a vaguely fish-like creature even more closely related to vertebrates than is *Ciona* — would be especially welcome.

What we have is a splendid start — or rather restart. This story really began in the 1880s with the frustration of one young researcher at the general inability to get to grips with the issue of vertebrate origins. With nothing to go on but basic embryology and comparative anatomy, schools of researchers devised a variety of explanations for vertebrate origins, not all of which were compatible, and with no way of deciding which ones were more likely to be correct.

This was the last straw for our hero. "Out of the same facts of anatomy and development men of equal ability and repute have brought the most opposite conclusions," he wrote. "To take for instance the question of the ancestry of the Chordata, the problem on which I was myself engaged, even if we neglect fanciful suggestions, there remain two wholly incompatible views as to the lines of Vertebrate descent, each well supported and upheld by many. From the same facts opposite conclusions are drawn. Facts of the same kind will take us no further. The issue turns not on the facts but on the assumptions."

He then proceeded to burn his boats in spectacular fashion: "Surely we can do better than this. Need we waste more effort in these vain and sophistical disputes?"

Six years after publishing these lines², the author coined a word for an entirely new discipline that promised to put hard facts in place of vacuous speculation. He was William Bateson, and the word was 'genetics'. Across the Atlantic, Thomas Hunt Morgan was also leaving old-fashioned evolutionary biology for experimental science, abandoning various worms for the fruitfly, on whose narrow and bristly shoulders rests a century

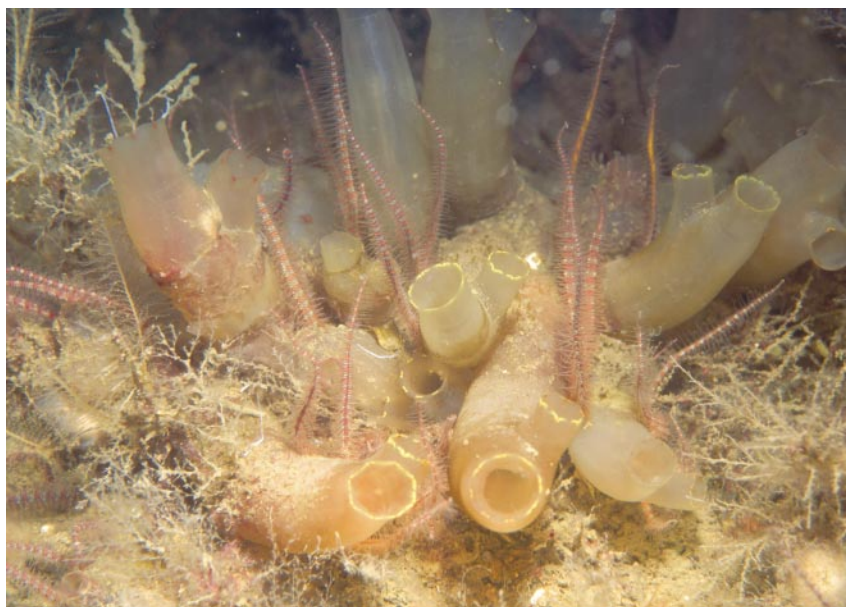


Figure 1 Object of taxonomic affection — *Ciona intestinalis*, the tube-like organisms seen in this underwater display.

of work on comparative developmental biology. The work on the genome of *Ciona* can thus trace its lineage to Bateson and Morgan and the very foundations of genetics.

The origin of vertebrates may seem like a question straight out of some dusty zoology textbook. But we owe the existence of genetics, genomics and, arguably, a large part

of molecular biology to the frustration of two young men unable to coax the secrets of evolution from creatures such as *Ciona intestinalis*. Welcome back, little squirt. ■
Henry Gee is a senior editor at Nature.

1. Dehal, P. et al. *Science* **298**, 2157–2167 (2002).

2. Bateson, W. *Materials for the Study of Variation* (Macmillan, London, 1894).

Computational biology

Evolution plays dice

Philip Gerrish

In computer simulations, initially identical populations of organisms growing in identical environments follow very different evolutionary trajectories. Mutational interdependence is a key factor.

Because of the long timescales involved, evolution has principally been a historical or retrospective science, relying in large part on a sparse fossil record. This constraint makes drawing inferences about evolutionary processes very difficult, but it does not apply to systems where evolution happens fast. With such systems, evolution ceases to be a strictly historical science and becomes amenable to experimentation, allowing hypotheses about evolution to be tested directly. Notably, fast evolution has been studied in laboratory populations of microbes^{1,2}, and really fast evolution has been studied in *in silico* populations of virtual organisms³. Virtual organisms are self-replicating, mutable computer programs, and as Yedid and Bell show on page 810 of this issue⁴, they provide a powerful way of addressing fundamental questions about evolution.

One such question, remarkably, continues to confound evolutionary biologists. Is evolution governed mainly by chance or is its outcome largely predetermined? With microbial or *in silico* systems, this question can be addressed experimentally as follows. Create several parallel populations with identical organisms; let them evolve independently in identical environments for a period of time; then see whether the evolved organisms are the same or not. If evolution is governed solely by chance, each population will follow a unique evolutionary trajectory and the evolved organisms will all be different. If evolution is completely deterministic, each population will follow the same trajectory and the evolved organisms will be the same.

This simple experiment has been performed with different microbes^{5–7}. Yet even with today's sophisticated biotechnological

tools, the level of observable detail is limited, restricting the inferences that can be made about the underlying processes. To look at those processes in more detail, Yedid and Bell⁴ performed the same experiment using populations of virtual organisms. Simply put, they conclude that chance rules.

This conclusion may seem obvious. After all, the mutations that drive evolutionary change occur at random, and any simulation that pretends to mimic them must do so using a random number generator. So, it seems, the accumulation of mutations over time must be a completely random process, whether *in vivo*, *in vitro* or *in silico*. The flaw in this logic is that it fails to acknowledge that a largely deterministic process — natural selection — is also at work in evolution.

Natural selection sorts out all the mutants in a population, sending the chaff to extinction and keeping only the best genotype — that is, the genetic constitution that confers the highest reproduction rate. In an unrealistically large (effectively infinite) population, all possible mutations are sampled, and so the best genotype will be predetermined by natural selection. In this extreme example, the outcome of evolution would be the same across replicate populations and one would expect the opposite of what Yedid and Bell found. But populations of realistic size are too small to sample all possible mutations (including double mutations, triple mutations, and so on). So the best genotype in one population will be different from the best genotype in another population. These two prizewinning genotypes may have similar reproduction rates, but their genetic configurations will probably be different. Indeed, Yedid and Bell found that their virtual bugs evolved similar biological properties (such as reproduction rate and genome size) across replicate populations, despite considerable divergence at the genetic level — a result that is corroborated by data from work with the bacterium *Escherichia coli*^{6,7}.

Evolutionary timescales are notoriously long, however. So it would seem that there ought to be plenty of time for each population to eventually acquire the best possible genetic configuration — which, again, should be the same across replicate populations. The reason this doesn't happen is that evolution occurs in 'steps', especially in asexual populations as studied by Yedid and Bell. Such steps happen when the best mutant available at the time becomes the predominant genotype in the population through the action of natural selection. At this point, the population is committed to carrying this particular mutation, even though it is probably not the best mutation the population could have acquired. That's because natural selection will not allow the population to go back to an inferior genotype. As if this commitment were not enough, the particular winning mutation largely determines which

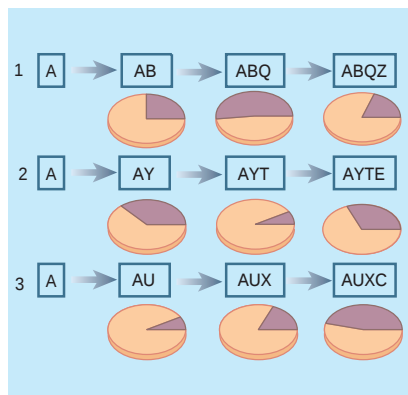


Figure 1 The population-divergence ratchet. All populations start with an identical genotype, A. Population 1 acquires mutation B, creating genotype AB. B is not the best possible mutation, but rather the best mutation then available. The pie chart indicates the increment in reproduction rate conferred by mutation B as a fraction (in this case, 75%) of the rate conferred by the best possible mutation. In population 3, mutation U confers 90% of the maximum possible increment. Because of epistasis, future evolutionary trajectories are restricted by previous mutations. For example, the B mutation is advantageous on the A background, as evidenced by its acquisition in population 1. But B may not be advantageous on the AU background, the second genotype in population 3. Instead, a different set of advantageous mutations is available to genotype AU. One of them is X, which is acquired, giving genotype AUX. From a new set of available mutations, C is then acquired. The same principles apply to population 2 and, overall, the ratchet produces ever-increasing population divergence, as observed in computer simulations by Yedid and Bell⁴.

of the subsequent mutations will further improve replication rate, one of which will cause the next evolutionary step. By carrying this particular winning mutation, therefore, the population is further committed to a new and limited set of future trajectories.

Such contingency on previous evolutionary steps results from the fact that the effects of mutations are interdependent. This phenomenon, known as epistasis, is often politely acknowledged to be quite prevalent in biological systems but is then promptly swept under the rug for the theoretical complications it causes. The results of Yedid and Bell's *in silico* 'go-back' experiments, however, suggest that ignoring epistasis in long-term evolution is a mistake. They find that interdependence among mutational effects — both physiological and social — plays a key role in the dynamics of population divergence. This idea is actually quite old⁸, but such strong experimental evidence supporting it is new and may be the most consequential contribution of Yedid and Bell's paper.

All in all, then, three factors together

produce a temporal asymmetry and thus a ratchet-like process that drives genetic divergence: the interdependence among mutational effects; the randomness associated with the best genotype available in a finite population; and the unidirectional tendency of natural selection to increase replication rate (Fig. 1). Not only did Yedid and Bell find considerable genetic divergence in their *in silico* populations, supporting the view that chance was the governing factor, but also the detail afforded by their experimental system allowed them to identify why. Populations diverge genetically because the role of chance is amplified by epistatic contingencies.

How much weight can we put on these conclusions? True, they do not come from real biological populations. But nor do they come from armchair speculation, a practice that has bedevilled evolutionary biology. Such experiments with virtual organisms are a step away from the armchair and towards experimental biology. But there is some way to go before virtual populations can be comparable to real biological populations. One potential difference is in population size. The largest virtual population size achieved by Yedid and Bell was 10^4 , whereas real populations of viruses, for example, can achieve sizes of the order of 10^{12} . Although the mutation supply rate can be adjusted in virtual populations to compensate for smaller population sizes (as Yedid and Bell did), the resulting population dynamics are not necessarily comparable to the dynamics of very large populations. As pointed out above, the degree of predictability in evolution should increase with increasing population size — a trend observed by Yedid and Bell and corroborated by data showing predictability in the evolution of HIV, where populations are typically very large⁹.

A second difference between real populations and virtual populations is that there is some degree of genetic exchange in most real populations, whereas Yedid and Bell's virtual populations were strictly asexual. Introducing genetic exchange may increase or decrease the degree of predictability in evolution — a question that could nicely be addressed with populations of virtual organisms. ■

Philip Gerrish is in the Programa de Investigación en Matemáticas Aplicadas y Computación, Instituto Mexicano del Petróleo, Col. San Bartolo Atepehuacán, México DF 07730, Mexico.
e-mail: pgerrish@yahoo.com

- Wahl, L. M. & Krakauer, D. C. *Genetics* **156**, 1437–1448 (2000).
- Appenzeller, T. *Science* **284**, 2108–2110 (1999).
- Zimmer, C. *Nat. Hist.* 42–45 (May 2001).
- Yedid, G. & Bell, G. *Nature* **420**, 810–812 (2002).
- Wichman, H. A., Badgett, M. R., Scott, L. A., Boulianne, C. M. & Bull, J. J. *Science* **285**, 422–424 (1999).
- Papadopoulos, D. et al. *Proc. Natl Acad. Sci. USA* **96**, 3807–3812 (1999).
- Lenski, R. E. & Travisano, M. *Proc. Natl Acad. Sci. USA* **91**, 6808–6814 (1994).
- Wright, S. *Genetics* **16**, 97–159 (1931).
- Boucher, C. A. et al. *J. Infect. Dis.* **165**, 105–110 (1992).

Obituary

René Thom (1923–2002)

When I first met René Thom, in 1970, he was already a legend in mathematics. There were many stories about him, which were reverently passed around. It was said, for instance, that his thesis adviser, back in the early 1950s, was concerned that he could not get Thom to write satisfactory mathematical proofs; another eminent mathematician told the adviser not to worry — there are ten people in the world, he said, who can prove these theorems once they have been stated, but only Thom can state them.

Indeed, Thom's own lectures gave some credence to such tales. He was a geometrician; he saw things in his mind and would try to impart his vision with all the means at his disposal, whether they were pictures on the blackboard or analogies from other sciences. In French mathematics, he was unique in this respect. There was general agreement at the time that serious mathematicians were in the business of devising proofs, not explaining them. The meaning of a theorem should be worked out from its formal proof. With Thom, it was the other way around. He showed why things were true, or should be true, and left his audience to work out the proofs.

Thom's research had much the same character: he inspired others, such as Bernard Malgrange or John Mather, to prove deep theorems that he alone, at the time, believed could be true. His own work has also deeply influenced modern mathematics. In 1958, he won the Fields medal — the equivalent in mathematics of the Nobel prize — for his early contributions to geometry.

But his most important work was still to come. I will try to explain it as Thom might have done, starting with the central idea from which everything else naturally grows. Thom asked, what is general, and what is singular? What does it mean for something to be ordinary, and when can we say that something is exceptional? An obvious answer to that question comes through probability theory: an event is general if there is a high probability that it will occur. We can say, for instance, that, in general, two straight lines in a plane will intersect: there is only one case in which they don't — when they are parallel. Similarly, two straight lines in three-dimensional space will, in general, not intersect.

Unfortunately, this probability analogy is too simple for many mathematical situations, but this was Thom's

achievement: he defined precisely what it means for a property to hold in general (he called such properties 'generic'). He devised a useful tool to check whether a given property is generic, known as Thom's transversality theorem (it sounds even better in French), and it is one of the workhorses of modern mathematics. Instead of looking for properties that are always true, now mathematicians very often try to find properties that hold in general, that are generic.

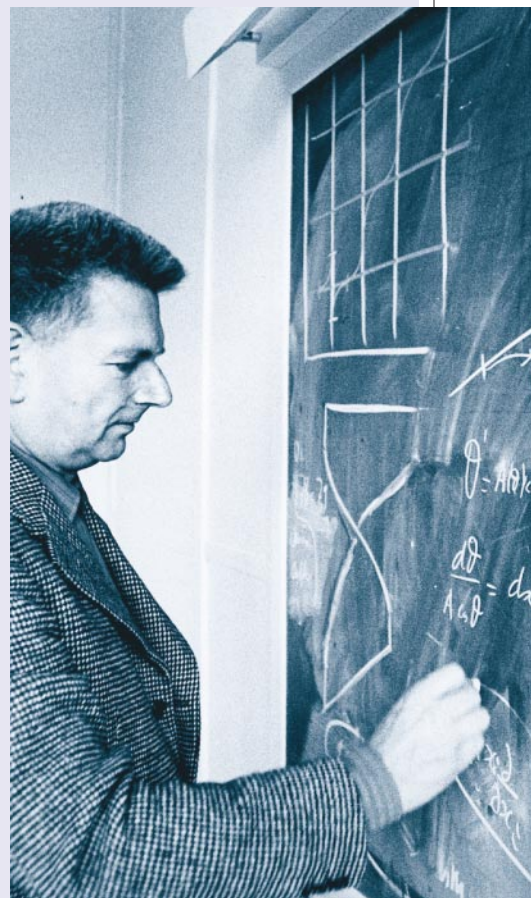
But Thom did not stop there. If a situation is exceptional, one may still ask how exceptional it is. To answer that question, a classification of singularities is needed, starting from the most frequent ones and working up to the less frequent (and presumably more complicated) ones. This was an enormous programme of research, and only a mind such as Thom's could have conceived it. The mathematics is difficult, and much of the work was done by John Mather. But it was successfully carried through for some

Mathematician who devised 'catastrophe theory'

particular cases, notably in the study of mappings between spaces of different dimensions, and the results are incredibly beautiful.

Thom called the first seven singularities the 'elementary catastrophes', and for a while pictures of the first three (the fold, the cusp and the swallowtail, which exist in three or fewer dimensions) were to be seen in many newspapers and magazines. Some claim that the hype was overdone, but the fact is that the underlying mathematics is very profound, and it is no mean achievement that Thom shared his vision with so many people who had no training in advanced mathematics.

The usefulness of this 'catastrophe theory', as Thom saw it, was in biology, in particular for morphogenesis — the generation of form and structure in an organism. At the time, it was not as fashionable for a mathematician to work on this kind of problem as it is now, but Thom went where his thoughts guided him. The sudden popularity of catastrophe theory took him down paths not usually trod by mathematicians, taking part (and holding his own) in numerous debates on science and philosophy. He also discovered a gift for writing, which he exercised



frequently, sometimes to the chagrin of his opponents.

From his PhD studies in Strasbourg in the early 1950s, Thom's career had taken him to Princeton, then back to France, to a position at Grenoble and a chair at Strasbourg. In 1961 he took up a professorship at the new Institut des Hautes Etudes Scientifiques, at Bures-sur-Yvette near Paris, where he remained until his retirement in 1988. He died on 25 October after a long illness, through which his family was a constant support.

Thom was not an aloof character, and he did not crave power. He built his career on his own merits. He was always accessible, always ready to listen. René Thom changed mathematics, just as a great painter changes our view of landscapes, and his vision is his legacy. But for those of us who knew him, there was more to it. This is best expressed by Heraclitus, in a sentence that Thom himself was fond of quoting: "The master whose oracle is in Delphi neither reveals nor hides, but gives hints."

Ivar Ekeland

Ivar Ekeland is in Cemerade et Institut de Finance, Université de Paris Dauphine, 75775 Paris, France.
e-mail: ekeland@dauphine.fr

A digital fluorescent molecular photoswitch

The signal from a gaudy ring molecule flashes on and off as light flicks it open and shut.

Fluorescent properties can be useful not only for tracking individual molecules within a microenvironment, but also in molecular-scale opto-electronics. Here we use external photostimulation to switch the fluorescence on and off from a single photochromic molecule embedded in a polymer film. This digital response is due to photoisomerization of the molecule, which may eventually find application in erasable optical data-storage elements.

Changes in fluorescence intensity from single molecules are not normally controllable, manifesting themselves as repeated cycles of emission ('blinking') or as spectral diffusion^{1–5}. Controlled on/off digital switching of chemically synthesized molecules by photo-irradiation has not yet been accomplished because of a lack of sufficiently robust molecules. The jellyfish green fluorescent protein (GFP) is the only example of a molecule that undergoes optically induced switching³ and has been exploited as a biological label in a cellular context.

Diarylethene derivatives are promising artificial photoswitching molecules because of their outstanding fatigue-resistant photochromic performance^{6–8}. After several trials⁹, we have designed and synthesized a fluorescent photoswitching molecule in which photochromic 1,2-bis(2-methoxy-4-phenyl-3-thienyl)perfluorocyclopentene and fluorescent 1,5-dimethoxy-9,10-bis(phenylethynyl)anthracene are linked through an adamantyl spacer (for structures, see supplementary information).

The open-ring form (**A**) of this molecule (see supplementary information) is converted to the closed-ring form (**B**) upon irradiation with ultraviolet light, and **B** returns to **A** under visible light; both forms are thermally stable⁶. The fluorescence spectrum of the bis(phenylethynyl)anthracene unit at 503 nm matches the absorption spectrum of the closed-ring form of the photochromic unit, so the fluorescence is efficiently quenched when the photochromic unit is converted from the open- to the closed-ring form upon ultraviolet irradiation.

The fluorescence quantum yields of **A** and **B** are 0.73 and <0.001, respectively. By measuring the fluorescence lifetimes ($\tau(\mathbf{A}) = 4.9$ ns, $\tau(\mathbf{B}) = 4.5$ ps), the intramolecular energy-transfer rate was calculated as $2.2 \times 10^{11} \text{ s}^{-1}$ and the energy-transfer efficiency as 99.9 %.

In the ensemble toluene solution, the fluorescence intensity gradually changes with the photo-isomerization from **A** to **B** and from **B** to **A**. The quantum yields of photocyclization (**A**→**B**) and photocyclo-

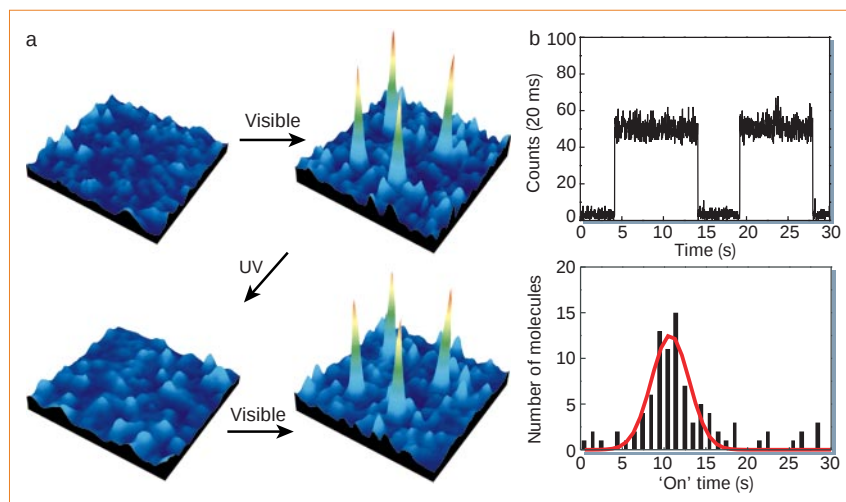


Figure 1 Fluorescence images and time traces of single photoswitching molecules of the open- and closed-ring forms of a compound in which a photochromic diarylethene derivative is linked to a fluorescent anthracene derivative through an adamantyl spacer. Photo-switching is detected by confocal microscopy. Samples were prepared by spin-coating a dilute toluene solution of the non-fluorescent, closed-ring form, **B** (2×10^{-11} M), on Zeonex thin polymer films (about 100 nm thick). **a**, Fluorescence images: top left, before irradiation; top right, after irradiation with visible light (wavelength, 488 nm; 200 W cm^{-2}) for 10 s; bottom left, after irradiation with weak ultraviolet light (325 nm, 0.027 mW cm^{-2}) for 3 s; bottom right, after irradiation with visible light (488 nm, 200 W cm^{-2}) for 10 s. **b**, Time trace of fluorescence intensity of a single molecule irradiated with strong 488-nm light (200 W cm^{-2}) and very weak 325-nm light (0.027 mW cm^{-2}), and histogram (below) of the time of switching on. The molecule was initially in the non-fluorescent **B** form. An accumulation time of 20 ms was chosen to avoid the effect of short-time blinking in the signal. All measurements were carried out under an atmosphere of nitrogen. We verified that single molecules were being detected by observing the fluorescence signal while rotating the polarization of linearly polarized laser light: the signal intensity was altered by the rotation and was reduced to the dark level.

reversion (**B**→**A**) in toluene were determined as 0.12 and 8.0×10^{-5} , respectively. The relatively large quantum yield of photocyclization is due to the very rapid cyclization reaction of the photochromic unit (<2 ps; ref. 10) and to the rigidity of the adamantyl spacer group.

The analogue-type photoresponse in the ensemble system changes to the digital-type one at the single-molecule level. Confocal microscopy reveals the fluorescence images and digital switching behaviour of single molecules embedded in a polymer film (Fig. 1a). Initially, the film containing non-fluorescent **B** molecules was dark, but when irradiated with visible light for 10 s, four molecules switched to the fluorescent 'on' state. Upon irradiation with weak ultraviolet light for 3 s, the fluorescent molecules then switched to the non-fluorescent 'off' state. The fluorescent state was restored by irradiating with visible light for 10 s. This digital reversible switching is due to photo-isomerization.

We investigated this switching in detail by measuring the response time of a single molecule to irradiation with strong visible and very weak ultraviolet light (Fig. 1b). Starting in the 'off' state, the non-fluorescent molecule abruptly switched to the fluorescent state at

4.1 and 19.2 s — these digital responses are due to photo-isomerization of **B** to **A**. At 14.1 and 28.0 s, the fluorescent molecule switched to the non-fluorescent state as a result of photo-isomerization of **A** to **B**.

To confirm that this switching is due to photo-isomerization, we measured the 'on' and 'off' times for 50–100 molecules by changing the power of the ultraviolet and visible light. The average 'on' and 'off' times shortened in proportion to the reciprocal power of the radiated light, indicating that the switching effect is indeed photochemical.

We have shown that digital switching of the fluorescence of diarylethene molecules can be controlled by irradiation with ultraviolet and visible light at the single-molecule level. Our fluorescent diarylethene derivative may find application in the design of erasable media for ultrahigh-density optical data storage.

Masahiro Irie*, **Tuyoshi Fukaminato***, **Takatoshi Sasaki***, **Naoto Tamai†**, **Tsuyoshi Kawai***

*Department of Chemistry and Biochemistry, Graduate School of Engineering, Kyushu University and CREST, Japan Science and Technology, Hakozaki 6-10-1, Higashi-ku, Fukuoka 812-8581, Japan
e-mail: irie@cstf.kyushu-u.ac.jp

†Department of Chemistry, School of Science and Technology, Kwansei Gakuin University, Gakuen 2-1, Sanda 669-1337, Japan

- Hofkens, J. et al. *J. Am. Chem. Soc.* **122**, 9278–9288 (2000).
- Lu, H. P. & Xie, X. S. *Nature* **385**, 143–146 (1997).
- Dickson, R. M., Cubitt, A. B., Tsien, R. Y. & Moerner, W. E. *Nature* **388**, 355–358 (1997).
- Ha, T., Enderle, T., Chempla, D. S., Selvin, P. R. & Weiss, S. *Chem. Phys. Lett.* **271**, 1–5 (1997).
- Nie, S., Chiu, D. T. & Zare, R. N. *Science* **266**, 1018–1021 (1994).

- Irie, M. *Chem. Rev.* **100**, 1685–1716 (2000).
- Hanazawa, M., Sumiya, R., Horikawa, Y. & Irie, M. *J. Chem. Soc. Chem. Commun.* 206–207 (1992).
- Irie, M., Kobatake, S. & Horichi, M. *Science* **291**, 1769–1772 (2001).
- Kawai, T., Sasaki, T. & Irie, M. *Chem. Commun.* 711–712 (2001).
- Tamai, N., Saika, T., Shimidzu, T. & Irie, M. *J. Phys. Chem.* **100**, 4689–4692 (1996).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Palaeontology

First Devonian tetrapod from Asia

The earliest tetrapods (vertebrates with limbs rather than paired fins) date from the Late Devonian Period (370–354 million years ago)^{1,2}—nine genera have been described, all of which are from the Euramerican supercontinent that comprises Europe, north America and Greenland, apart from a single Gondwanan genus, *Metaxygnathus*, from Australia^{3–5}. Here we report the discovery of the first Devonian tetrapod from Asia, a finding that substantially extends the geographical range of these animals and raises new questions about their dispersal. These forms seem to have achieved worldwide distribution and great taxonomic diversity within a relatively short time.

The new fossil, an incomplete left mandible (Fig. 1a; Box 1), was collected by us from the Late Devonian (about 355 million years BP) non-marine Zhongning Formation of the Ningxia Hui autonomous region, northwestern China, in July 2002. Associated fossils include the plants *Lepto-*

phloeum rhombicum and *Sublepidodendron mirabile*, land plant spores, the antiarch fishes *Remigolepis* and *Sinolepis*⁶, and undescribed lobe-finned fish remains. The specimen is exposed in mesial view and comprises most of the prearticular, together with the angular and postsplenial (exposed along the ventral jaw margin). The dentary and coronoids are missing.

Mandibles of early tetrapods differ substantially from those of lobe-finned fish⁴, allowing this specimen to be identified with confidence as a tetrapod. The main diagnostic characters are the appearance of the prearticular, which carries a well-defined dorsal band of denticles but is otherwise smooth apart from a faint radial striation, and the absence of Meckelian ossification between the prearticular and infradentaries. Lobe-finned fish have denticles that extend across most or all of the prearticular, and a well-ossified Meckelian element that binds the prearticular and infradentaries together. The absence of the dentary is also significant: this bone is loosely attached in many Devonian tetrapods⁴ and is often lost from isolated jaws⁵, but in lobe-finned fish it is firmly sutured to the jaw and does not detach.

Box 1 Taxonomic description

Tetrapoda (Goodrich, 1930)

Sinostega pani Zhu and Ahlberg gen. et sp. nov.

Provisional diagnosis. A stem-group tetrapod lacking Meckelian ossification between the prearticular and infradentaries (angular and postsplenial) in the middle part of the lower jaw. The denticulated field on the prearticular does not reach as far anteriorly as in *Acanthostega*.

Etymology. Generic name from Latin *Sino-* (pertaining to China) and Greek *stega* (roof, now a conventional ending for Devonian tetrapod names, cf. *Ichthyostega*, *Ventastega*). Specific name in honour of Pan Jiang as contributor to the study of the Ningxia Devonian biota⁶.

Holotype. V13576, an incomplete lower-jaw ramus. IVPP, Beijing.

Locality and horizon. Ningxia Hui autonomous region, China (see Fig. 1c). Zhongning Formation (late Famennian, Late Devonian).

Remarks. The absence of Meckelian ossification in the middle part of the jaw distinguishes *Sinostega* from all known Devonian tetrapods except *Acanthostega* and possibly *Tulerpeton*. It most closely resembles *Acanthostega*⁴, but differs from that genus in having a smaller denticulated field on the prearticular.

During the Late Devonian, northern China lay in tropical latitudes adjacent to northeastern Gondwana (Fig. 1b)⁷. It is therefore interesting that *Sinostega* resembles the Euramerican genus *Acanthostega* more closely than it does the Australian *Metaxygnathus*. Tetrapods probably originated in Euramerica during the Frasnian Age^{4,8,9}. We consider that tetrapods had achieved an essentially global distribution in tropical to subtropical latitudes by the end of the Famennian, some 5–10 million years later (supporting the idea that the earliest tetrapods might have been littoral¹⁰), and that their apparent scarcity in northern Gondwana and China is a sampling artefact that will be remedied by future discoveries.

Min Zhu*, Per E. Ahlberg†, Wenjin Zhao*, Liantao Jia*

*Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, PO Box 643, Beijing 100044, China
e-mail: zhumin@ivpp.ac.cn

†Department of Palaeontology, Natural History Museum, London SW7 5BD, UK

- Ahlberg, P. E. & Milner, A. R. *Nature* **368**, 507–514 (1994).
- Coates, M. I. & Clack, J. A. *Nature* **352**, 234–236 (1991).
- Campbell, K. S. W. & Bell, M. *Alcheringa* **1**, 369–381 (1977).
- Ahlberg, P. E. & Clack, J. A. *Trans. R. Soc. Edinb.* **89**, 11–46 (1998).
- Daeschler, E. B. *J. Paleontol.* **74**, 301–308 (2000).
- Pan, J. et al. *Continental Devonian System of Ningxia and its Biota* (Geological, Beijing, 1987).
- Metcalfe, I. in *Biogeography and Geological Evolution of SE Asia* (eds Hall, R. & Holloway, J. D.) 25–41 (Backhuys, Leiden, The Netherlands, 1998).
- Ahlberg, P. E. *Nature* **373**, 420–425 (1995).
- Ahlberg, P. E., Lukševičs, E. & Mark-Kurik, E. *Palaeontology* **43**,

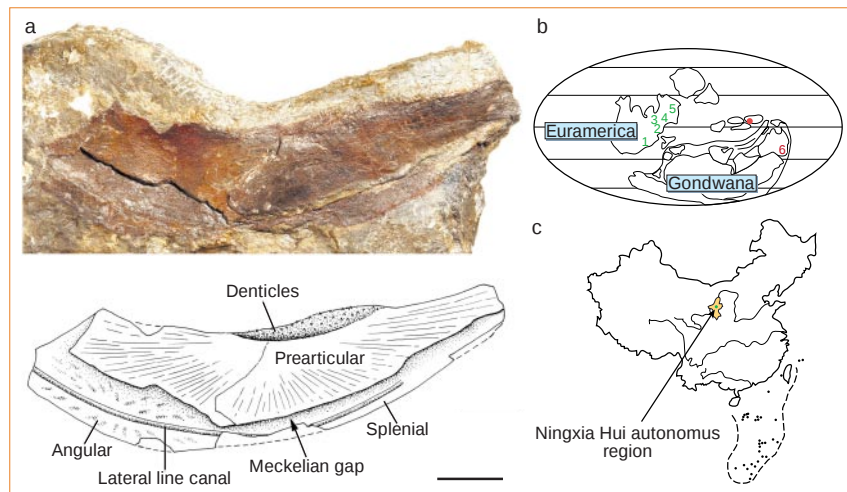


Figure 1 A new Devonian tetrapod from China. **a**, Photograph (top) and interpretative drawing (bottom) of IVPP V13576, a lower-jaw ramus from the Ningxia Hui autonomous region, China. Note the absence of Meckelian ossification (Meckelian gap) between the prearticular and infradentaries (splenial and angular), and the narrow denticulated band along the top of the prearticular, both of which are diagnostic tetrapod features. Scale bar, 10 mm. **b**, Late Devonian palaeogeographical map (modified from ref. 7) showing the location of the new Devonian tetrapod (red dot) in relation to previously discovered specimens. Green numbering, Euramerican taxa; red, non-Euramerican: 1, Pennsylvania (*Hynerpeton*, *Densignathus*); 2, Scotland (*Elginerpeton*); 3, Greenland (*Ichthyostega*, *Acanthostega*); 4, Latvia (*Obruchevichthys*, *Ventastega*); 5, Russia (*Tulerpeton*); 6, New South Wales (*Metaxygnathus*). **c**, Map of modern-day China, showing the Ningxia Hui autonomous region in orange; green dot indicates the fossil site, close to the Yellow River.

533–548 (2000).

10. Milner, A. R. in *Palaeozoic Vertebrate Biostratigraphy and Biogeography* (ed. Long, J. A.) 324–353 (Belhaven, London, 1993).

Competing financial interests: declared none.

Nanotechnology

Carbon nanotubes with DNA recognition

Since the discovery of their one-dimensional electronic band structure¹, the leading candidate that has emerged for nanodevice applications is single-walled carbon nanotubes (SWNTs). Here we unite their unique properties with the specific molecular-recognition features of DNA by coupling SWNTs to peptide nucleic acid (PNA, an uncharged DNA analogue²) and hybridizing these macromolecular wires with complementary DNA. Our findings provide a new, versatile means of incorporating SWNTs into larger electronic devices by recognition-based assembly, and of using SWNTs as probes in biological systems by sequence-specific attachment.

At present, SWNT-based devices such as field-effect transistors³, logic circuits^{4,5} and single-electron transistors⁶ are fabricated by 'top-down' lithographic methods. The construction of more complex architectures with high device density requires the development of a 'bottom-up', massively parallel strategy that exploits the molecular properties of SWNTs.

We have therefore developed a technique to couple SWNTs covalently to PNA. Our process begins by ultrasonically shortening

SWNT ropes (Tubes@Rice) for 1 hour in a 3:1 mixture of concentrated H₂SO₄ and HNO₃. Subsequent exposure to 1 M HCl produces abundant carboxyl end-groups⁷. This material is dispersed in dimethylformamide (DMF, 99.5%) and incubated for 30 min in 2 mM 1-ethyl-3-(3-dimethylaminopropyl)carbodiimide hydrochloride and 5 mM *N*-hydroxysuccinimide (NHS) to form SWNT-bearing NHS esters⁸ (Fig. 1a).

PNA adducts are formed (Fig. 1b) by reacting this material in DMF for 1 hour with excess PNA (sequence: NH₂-Glu-GTGCTCATGGT-CONH₂, where Glu is a glutamate amino-acid residue and the central block represents nucleic-acid bases; Fig. 1b). The PNA-derivatized SWNTs are transferred to water and dispersed in 0.5% aqueous sodium dodecyl sulphate, which stabilizes the individual SWNTs⁹.

To investigate whether DNA would hybridize to PNA-SWNTs, we prepared fragments of double-stranded DNA with 12-base-pair, single-stranded 'sticky' ends that were complementary to the PNA sequence. These fragments were produced by cutting double-stranded DNA with the restriction enzyme *Hind*III and ligating the products to single-stranded oligonucleotides. This sticky DNA was hybridized to the PNA-SWNTs (Fig. 1c) in water, deposited on freshly cleaved mica with 5 mM MgCl₂, and the surface was rinsed and dried after 30 s. Atomic-force micrographs (Fig. 1d, e) of the DNA/PNA-SWNT hybrids were recorded under ambient conditions.

From our observations of several samples, we conclude that DNA attachment

occurs predominantly at or near the nanotube ends. Because the derivatization chemistry is done on SWNT 'ropes', we contend that individual SWNTs are largely shielded, resulting in a low density of defects in their side walls. The rare attachment of DNA to other regions of SWNTs also indicates that this is the result of sequence-specific PNA–DNA base-pairing, rather than of nonspecific interaction.

We chose to couple PNA, rather than DNA, directly with SWNTs for several reasons¹⁰. First, PNA is compatible with the most convenient solvents (DMF, for example); second, PNA is not susceptible to enzymatic degradation; and third, the uncharged PNA backbone gives rise to PNA–DNA duplexes that are more thermally stable than their DNA–DNA counterparts because there is no electrostatic repulsion. Besides reducing the length of the sticky ends required for room-temperature hybridization, this last property should reduce nonspecific electrostatic interactions with metallic electrodes or with surfaces, such as silicon oxide, that are convenient for lithography.

The recognition properties imparted to SWNTs by oligonucleotide adducts could be used to programme the attachment of SWNTs to each other and to substrate features, such as electrodes, on which monolayers of complementary sequences are self-assembled. The antisense properties of PNA-SWNTs might also be exploited in a biological context, for example in biosensors. Our results represent a step towards full compatibility of SWNTs with enzymes and proteins^{11,12} — a powerful approach for organizing complex devices at the sublithographic scale.

Keith A. Williams*, Peter T. M. Veenhuizen*, Beatriz G. de la Torre†, Ramon Eritja†, Cees Dekker*

*Department of NanoScience, Delft University of Technology, 2628 CJ Delft, The Netherlands
e-mail: williams@mb.tn.tudelft.nl

†Instituto de Biología Molecular Barcelona, Consejo Superior de Investigaciones Científicas, 08034 Barcelona, Spain

1. Mintmire, J. W., Dunlap, B. I. & White, C. T. *Phys. Rev. Lett.* **68**, 631–634 (1992).
2. Nielsen, P. E., Engholm, M., Berg, R. H. & Buchardt, O. *Science* **254**, 1497–1500 (1991).
3. Tans, S., Verschueren, A. R. M. & Dekker, C. *Nature* **393**, 49–52 (1998).
4. Derycke, V., Martel, R., Appenzeller, J. & Avouris, P. *Nano Lett.* **1**, 453–456 (2001).
5. Bachtold, A., Hadley, P., Nakanishi, T. & Dekker, C. *Science* **294**, 1317–1320 (2001).
6. Postma, H. W. C., Teepen, T. F., Yao, Z., Grifoni, M. & Dekker, C. *Science* **293**, 76–79 (2001).
7. Chen, J. *et al. Science* **282**, 95–98 (1998).
8. Grabarek, Z. & Gergely, J. *Analyt. Biochem.* **185**, 131–135 (1990).
9. O'Connell, M. J. *et al. Science* **297**, 593–596 (2002).
10. Wang, J. *Biosensors Bioelectron.* **13**, 757–762 (1998).
11. Wong, S. S., Joselevich, E., Woolley, A. T., Cheung, C. L. & Lieber, C. M. *Nature* **394**, 52–55 (1998).
12. Huang, W. *et al. Nano Lett.* **2**, 311–314 (2002).

Competing financial interests: declared none.

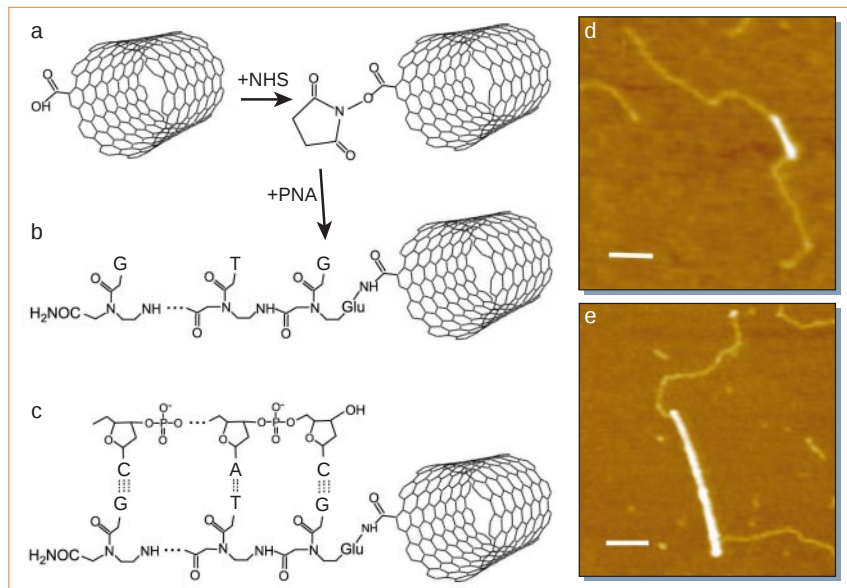


Figure 1 Attachment of DNA to carbon nanotubes. **a**, **b**, *N*-hydroxysuccinimide (NHS) esters formed on carboxylated, single-walled carbon nanotubes (SWNTs) are displaced by peptide nucleic acid (PNA), forming an amide linkage. **c**, A DNA fragment with a single-stranded, 'sticky' end hybridizes by Watson–Crick base-pairing to the PNA-SWNT. **d**, **e**, Atomic-force microscope (TappingMode) images of PNA-SWNTs. SWNTs appear as bright lines; the paler strands represent bound DNA. Scale bars: 100 nm; nanotube diameters: **d**, 0.9 nm; **e**, 1.6 nm.

Secure communication

Quantum cryptography with a photon turnstile

Quantum cryptography generates unbreakable cryptographic codes by encoding information using single photons, which until now have relied on highly attenuated lasers as sources^{1,2}. But these sources can create pulses that contain more than one photon, making them vulnerable to eavesdropping by photon splitting^{3,4}. Here we present an experimental demonstration of quantum cryptography that uses a photon turnstile device, which is more reliable for delivering photons one at a time. This device allows completely secure communication in circumstances under which this would be impossible with an attenuated laser.

Our quantum-cryptography system (see supplementary information for full technical details) implements a protocol known as BB84 (ref. 5). The photon turnstile is a single quantum dot in a micropost cavity^{6,7}, which is optically excited by a pulsed laser. The security improvements attainable with this device can be quantified by two measurements: the probability that the device will inject a photon into the quantum channel, measured as 0.007 by comparing the count rate at detector 0 (see supplementary information) to the repetition rate of the excitation laser (76 MHz); and the second-order correlation, denoted by $g^{(2)}$ (see supplementary information).

This quantity gives the amount of suppression of multiphoton states from our device relative to attenuated laser light — a laser with perfect intensity stability is characterized by $g^{(2)} = 1$, whereas our turnstile device has $g^{(2)} = 0.14$. The probability

that our device will emit a multiphoton state is therefore an order of magnitude smaller than a laser that emits photons at the same rate, meaning that security is improved in the presence of channel losses⁸.

In our implementation of BB84, the sender of the message, Alice, encodes information by preparing the polarization of each photon in either the horizontal or right circular polarization for binary 0, and vertical or left circular for binary 1. This is done by an electro-optic modulator. The modulator is driven by a data generator that produces the secret key, giving a random four-level signal that corresponds to the four different polarization states in the BB84 protocol. The state of the data generator is recorded by a time-interval analyser and is stored by a computer.

After the polarization is prepared, the photon is sent into the quantum channel, a 1-metre free-space propagation, and is detected by the receiving party, Bob. Bob measures the photons by using passive polarization optics and avalanche photodiodes with dark counts of about 80 s^{-1} (see supplementary information). The detection probability, due to losses in the optics and photodiodes, is 0.24. Detection events are recorded by a second time-interval analyser and are stored by a second computer for subsequent comparison with Alice.

The error rate of the system is measured as 2.5%. These errors are corrected by using an error-correction algorithm⁹. After error correction, privacy amplification is carried out to create the final key, yielding a communication rate of 25 kbits s^{-1} .

To demonstrate the advantage of our source over standard laser light, we use both the turnstile device and an attenuated laser to carry out quantum cryptography. A variable attenuator is inserted into the quantum channel to simulate channel losses. Figure 1 shows the theoretical and experimentally measured communication rates for our turnstile device and for laser light, as a function of channel loss. When losses are low, the communication rate of the attenuated laser is greater because our turnstile device is limited by its efficiency and by losses in the collection optics. At greater channel losses, however, the laser emits too many multiphoton states, causing a more rapid reduction in communication rate. Above 23 dB of loss, secure communication is no longer possible with the laser — however, our source is able to withstand channel losses of about 28 dB. This demonstrates the security advantage of our photon turnstile in the presence of channel losses.

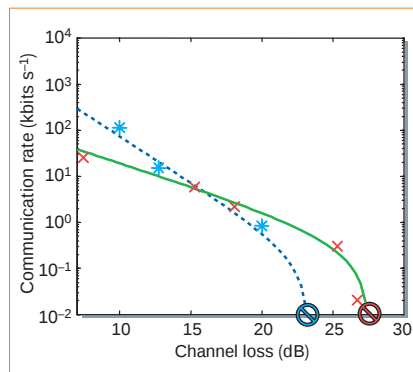


Figure 1 Comparison of the turnstile device with a standard laser. Measured (crosses, turnstile device; stars, laser) and simulated (full line, turnstile device; dashed line, laser) bit rates are shown as a function of total loss from the channel and detection system. Circles: blue, the attenuation at which our system was experimentally shown to reject the entire key using an attenuated laser (23 decibels); red, point at which our system rejected the entire key for the turnstile device (28 decibels). This shows a 5-decibel improvement in the loss cut-off.

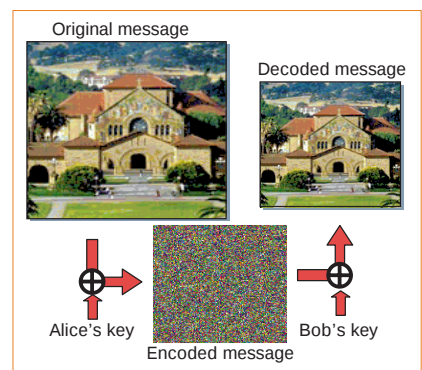


Figure 2 Demonstration of the final stage of the cryptographic protocol. The message, a $140 \times 141,256$ -pixel colour bitmap of Stanford University's Memorial Church, is encoded by a secure quantum key exchanged through our system from Alice to Bob, using standard one-time pad techniques. The encoded image appears as white noise to a third party that does not possess a copy of the key. Using the exchanged key, however, Bob decodes the message, recovering the pixel image with no error. Further details are available from the authors.

In the final phase of communication, the secret key is used as a one-time pad to exchange the message (here a picture of Stanford University's Memorial Church; Fig. 2). The cryptography system is used to exchange a 20-kilobyte key. Alice uses her copy of the key to do a bitwise exclusive OR logic operation with each bit of the message. The resulting encrypted message looks like white noise to anyone without a copy of the key, but Bob decodes it by carrying out a second bitwise exclusive OR operation using his copy of the key.

A similar experiment using diamond colour centres has recently been reported¹⁰.

Edo Waks*, Kyo Inoue†, Charles Santori*, David Fattal*, Jelena Vuckovic*, Glenn S. Solomon‡, Yoshihisa Yamamoto*†

*Quantum Entanglement Project, ICORP, JST, MURI, UCLA, E. L. Ginzton Labs, and ‡Solid State Photonics Laboratory, Stanford University, Stanford, California 94305, USA

e-mail: edo@stanford.edu

†NTT Basic Research Laboratories, Atsugi, Kanagawa 243-0198, Japan

- Buttler, W. T. et al. *Phys. Rev. Lett.* **24**, 5652–5655 (2000).
- Marand, C. & Townsend, P. D. *Optics Lett.* **20**, 1695–1697 (1995).
- Brassard, G., Lütkenhaus, N., Mor, T. & Sanders, B. C. *Phys. Rev. Lett.* **85**, 1330–1333 (2000).
- Lütkenhaus, N. *Phys. Rev. A* **61**, 052304 (2000).
- Bennett, C. H. & Brassard, G. *Proc. IEEE Int. Conf. Comp. Syst. Sig. Process., Bangalore, India* 175–179 (IEEE, New York, 1984).
- Santori, C., Pelton, M., Solomon, G., Dale, Y. & Yamamoto, Y. *Phys. Rev. Lett.* **86**, 1502–1505 (2001).
- Santori, C., Fattal, D., Vuckovic, J., Solomon, G. S. & Yamamoto, Y. *Nature* **419**, 594–597 (2002).
- Waks, E., Santori, C. & Yamamoto, Y. *Phys. Rev. A* **66**, 042315 (2002).
- Brassard, G. & Salvail, L. *Advances in Cryptology: Eurocrypt '93, Lecture Notes in Computer Science Vol. 765* (ed. Hellseth, T.) 410–423 (Springer, Berlin, 1994).
- Beveratos, A. et al. *Phys. Rev. Lett.* **89**, 187904 (2002).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.

Measurement of polarization with the Degree Angular Scale Interferometer

E. M. Leitch^{*†‡}, J. M. Kovac^{§†‡}, C. Pryke^{*†‡||}, J. E. Carlstrom^{*§†‡||}, N. W. Halverson^{¶†}, W. L. Holzapfel^{¶†}, M. Dragovan^{*†||}, B. Reddall[†] & E. S. Sandberg[†]

^{*} Department of Astronomy & Astrophysics; [†] Center for Astrophysical Research in Antarctica; [‡] Center for Cosmological Physics; [§] Department of Physics; and ^{||} Enrico Fermi Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA

[¶] Department of Physics, University of California at Berkeley, Le Conte Hall, California, 94720, USA

Measurements of the cosmic microwave background (CMB) radiation can reveal with remarkable precision the conditions of the Universe when it was $\sim 400,000$ years old. The three most fundamental properties of the CMB are its frequency spectrum (which determines the temperature), and the fluctuations in both the temperature and polarization across a range of angular scales. The frequency spectrum has been well determined, and considerable progress has been made in measuring the power spectrum of the temperature fluctuations. But despite many efforts to measure the polarization, detection of this property of the CMB has hitherto been beyond the reach of even the most sensitive observations. Here we describe the Degree Angular Scale Interferometer (DASI), an array of radio telescopes, which for the past two years has conducted polarization-sensitive observations of the CMB from the Amundsen–Scott South Pole research station.

DASI is an interferometric array designed to measure anisotropy in the CMB radiation on angular scales of $1.3\text{--}0.2^\circ$, corresponding to multipoles $l \approx 140\text{--}900$. Arrays of telescopes operate as interferometers by measuring the correlations of the signals received by pairs of telescopes. In a radio interferometer such as DASI, the correlations are formed by taking the time average of the product of the signals. The effective response on the sky for each pair of telescopes is similar to a two-slit interference pattern; in the limit of infinitesimally small telescopes, the pattern would be a simple sinusoidal corrugation. In this case, it is easy to see that an interferometer measures the Fourier components of the sky brightness distribution. The finite size of the telescopes sets the field of view of the interferometer by multiplying the corrugation with the beam pattern of the individual telescopes. An interferometer with finite apertures thus measures a finite range of Fourier components.

The ability of interferometers to make simultaneous differential measurements of the sky brightness, that is, directly measure Fourier components, gives them enormous stability and makes them an attractive choice for measuring the angular power spectra of the CMB. Traditional interferometers, however, have been designed for high angular resolution, beyond what could be provided by a single large-diameter telescope. Such arrays consist of large individual telescopes, yielding high sensitivity to a small region of the sky. The Very Large Array, comprising 25-m telescopes distributed over tens of kilometres, is one such example. By contrast, the scales of interest for CMB anisotropy measurements are of the order of a degree, requiring telescope separations of mere tens of wavelengths. DASI, which operates at centimetre wavelengths, consists of thirteen 20-cm-diameter telescopes with separations spanning 25–121 cm.

The DASI telescope was deployed at the Amundsen–Scott South Pole Station in the 1999–2000 austral summer, and made total intensity measurements of the CMB during the 2000 austral winter. Details of the instrument, the measured power spectrum and the resulting cosmological constraints were presented in a series of three papers^{1–3}.

During the 2000–01 austral summer, the DASI receivers were each fitted with broadband achromatic polarizers to allow polarization-sensitive observations of the CMB. In addition, a large reflecting ground screen was installed to reduce the sensitivity to terrestrial sources of emission. Throughout the 2001 and 2002 austral winters, the telescope has observed the CMB in all four Stokes parameters.

Here we describe the design of the DASI CMB polarization experiment, the polarization response and calibration of the instrument, and the CMB observations made during the 2001 and 2002 seasons. The analysis of the CMB polarization data obtained is presented in a companion paper⁴.

Description of the instrument

DASI consists of an array of 13 lensed, corrugated feed horns, 20 cm in diameter, each surrounded by a corrugated shroud to suppress cross-talk between the horns. Signals from the array are correlated in ten 1-GHz bands over the frequency range 26–36 GHz, with horn separations ranging from 25–121 cm, sampling points in the Fourier (u, v) plane at radii of $22\text{--}143 \lambda^{-1}$, or equivalently, multipoles in the range $l \approx 140\text{--}900$ at the frequency extrema.

The DASI feed horns are distributed in a three-fold symmetric pattern on a rigid faceplate, which can be rotated about its axis to observe at a range of parallactic angles. The entire faceplate is in turn attached to an altitude–azimuth mount. DASI and its companion instrument, the Cosmic Background Imager⁵, differ from conventional interferometers in that they are co-planar arrays, and the projected baselines do not change as a source is tracked around the sky. This allows arbitrarily long integration on a given baseline with no smearing in the Fourier plane, and permits a variety of consistency checks; given the symmetry of DASI horn configuration, rotations of the faceplate that are integral multiples of 60° produce identical Fourier-space sampling.

Receiver polarization hardware

The first stage of each DASI receiver is a single-mode (that is, a single polarization state) HEMT (high electron mobility transistor) amplifier. DASI does not use orthomode transformers to separate orthogonal polarization states. Instead, a mechanically switchable waveguide polarizer is inserted between the amplifier and the feedhorn to select between left- and right-handed polarization states, L and R.

The DASI horns feed circular waveguide, which by its symmetry supports two polarization modes. The circular waveguide is gradually tapered to single-mode rectangular waveguide that supports only one mode of linear polarization. For sensitivity to circular polarization, a transformation from circular to linear polarization must therefore be made in the circular waveguide or outside the

feed. The standard technique uses a $\lambda/4$ retarder oriented 45° to the E -field of the linear polarization accepted by the rectangular waveguide. This retarder is usually a simple dielectric vane inserted in the circular waveguide; at 30 GHz, the loss in the dielectric and the associated noise added to the system are negligible if the vane is cooled. The circular waveguide that holds the vane can be mechanically rotated along the waveguide axis by $\pm 45^\circ$ to select L or R.

DASI uses a mechanically switched circular waveguide dielectric polarizer similar to that outlined above. A simple dielectric, however, retards by precisely $\lambda/4$ at only a single frequency; at frequencies offset from this design frequency, deviation from the $\lambda/4$ behaviour leads to a contribution from the unwanted circular polarization state, known as 'leakage'. Owing to DASI's large fractional bandwidth (30%), the leakage for a simple $\lambda/4$ retarder polarizer would be prohibitively large at the band edges (see Fig. 1).

To construct broadband, low-leakage polarizers, we investigated designs using multiple retarder elements and found that in principle there are solutions to meet any specified polarization purity and bandwidth. Details of the theory of multi-element achromatic waveguide polarizers and details of the DASI polarizers, including extensive laboratory test results, will be presented elsewhere (J. M. K. and J. E. C., manuscript in preparation). For the DASI polarizers we selected a two-element design composed of $\lambda/2$ and $\lambda/4$ dielectric retarding waveguide vanes made from 1.3-mm-thick polystyrene. With a two-element design, the retarders can be arranged so that the offset-frequency errors cancel to first order. The design also incorporates an absorbing vane to suppress the linear polarized mode that is reflected by the circular to rectangular waveguide transition.

To aid in understanding the DASI design, consider the signal path as if the receivers were in fact transmitters starting with linear polarization in the rectangular waveguide which the polarizer transforms to circular polarization. In the DASI design, the $\lambda/2$ retarder simply introduces a rotation of the E -field of the linear polarization at the design frequency, but also introduces offset frequency errors that increase towards the band edges. The $\lambda/4$ retarder is oriented 45° to the rotated E -field to produce the desired circular polarization. The orientation of the $\lambda/2$ retarder to the incident polarization is chosen such that dispersion effects of the two retarders cancel to first order. The entire assembly is then mechanically rotated to select L or R polarization. The angle between the $\lambda/2$ and $\lambda/4$ retarders is held fixed at 60° . The angle of the $\lambda/2$ retarder to the E -field of the linearly polarized wave switches between 15° and 105° to produce the orthogonal circular polarization states.

In Fig. 1, we show the predicted leakage for the DASI two-element design and, for comparison, the traditional single $\lambda/4$ design. Also

shown in Fig. 1 is the leakage response determined from astronomical observations, as discussed in the subsection 'On-axis leakage', below.

The stability of the polarization leakages is more critical than their magnitude, because the leakages, if stable, can be corrected for in the data analysis. For DASI the stability of the polarizers is determined by the repeatability and accuracy with which the angle is set. The polarizers are driven by a stepper motor with a 0.9° half-step, geared down by a factor of three using anti-backlash gears, to provide 0.30° steps of the vane assembly. The position of the polarizer is read by an encoder with 0.12° resolution. All of the hardware is contained under vacuum within the receiver cryostats. The motor and encoder are warm, but the waveguide vane assembly is cooled to ~ 10 K. With 2.5 counts of the polarizer encoder for the minimum step of the motor, we can ensure a repeatable return to the same position. We have verified that the stepper motor holds the polarizer position to better than half of an encoder count ($<0.06^\circ$) for the entire season. This accuracy translates to a variation in the leakage of less than 0.1%. Astronomical tests of the stability of the leakages on long timescales are given in the subsection 'On-axis leakage', below.

Ground shield

A ground shield was installed around the telescope during the 2000–01 austral summer, and was in place for the entirety of the 2001–02 observations presented here. The ground shield is designed to reflect the antenna sidelobes onto the cold sky, and consists of three concentric rings of 12 panels, the innermost flat, the outer two rising at a pitch that increases from $\sim 30^\circ$ for the inner ring to $\sim 60^\circ$ for the outer (see Fig. 2). The design of the panels was guided by concerns about crosstalk between close-packed array elements; panels are configured to prevent sidelobes up to 90° off-axis from seeing any other horns in reflection.

The top of the outermost ring extends to a height of 13 feet (1 ft = 0.3048 m) above the roofline. The uppermost border of the shield is rolled with a 1-ft-radius to reduce diffraction around the edge, and all gaps between panels are covered with aluminium adhesive tape. Four of the upper panels are hinged and may be lowered to allow observations of planets, or a test transmitter mounted on the roof of one of the station buildings. Without lowering the shields, the minimum observing elevation is $\sim 25^\circ$.

Although the DASI telescope was intended to operate with the ground shield from its inception, the shield panels were not in place during the year-2000 observations reported in refs 1 and 2, and considerable effort was devoted to reducing the effect of the residual ground signal in that data set. Throughout those observations, the effect of the ground was easily visible in the raw data and was a strong function of (u, v) radius.

During both 2000 and 2001, fields were observed in consecutive 1-h blocks over the same azimuth range (see the section 'CMB observations', below), and we can probe for residual contamination from the ground by forming the difference between successive 1-h

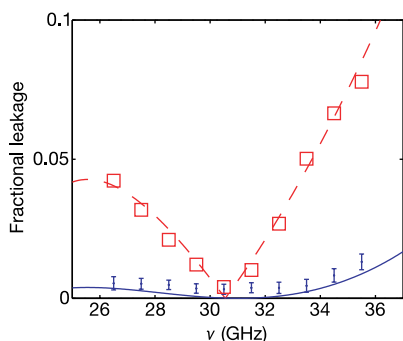


Figure 1 Comparison of the instrumental polarization for single and multi-element polarizers. Shown are the theoretical minimum instrumental polarization for a single-element polarizer (dashed line), and for a two-element polarizer (solid line). Squares are bench-top measurements of a single-element polarizer. Points are the astronomical measurements of the average on-axis instrumental polarization for the 13 DASI receivers configured with two-element polarizers.

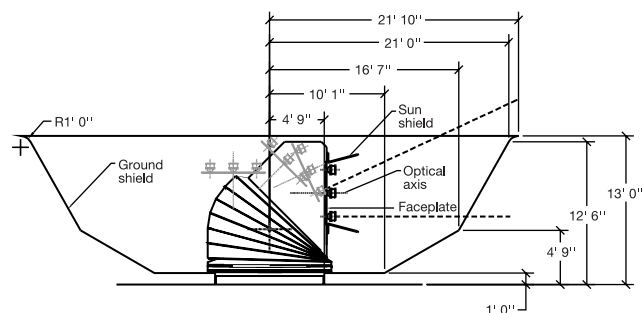


Figure 2 Ground shield geometry. The ground shields are designed to reflect sidelobes of the horns to the cold sky, and to prevent sidelobes from seeing other horns, owing to concerns about crosstalk.

means. In the 2000 data, taken without the ground shield, comparison of the variance in these 2-h differences to the variance in data differenced at the 8-s integration rate indicates a significant (5%) noise excess, dominated by the data on the shortest baselines. In the 2001 co-polar data, taken with the ground shield in place, the noise excess has dropped by nearly an order of magnitude. Any noise excess in the cross-polar data is a factor of 3–4 lower still, indicating that any residual signal from the ground is largely unpolarized. See the ‘Polarization response’ section for a definition of co-/cross-polar.

Sun shield

When the Sun is above the horizon, data taken over much of the sky is contaminated by signal entering the far-off sidelobes of the horn beam pattern. This contamination is most evident as long-period fringing in visibility data taken within $\sim 60^\circ$ of the Sun (for a definition of visibility, see the subsection ‘Polarization response’, below). CMB data are consequently collected only during the winter, when the Sun is below the horizon, while the summer season (when the station is accessible) is reserved for maintenance of the telescope. Although sunset and sunrise at the pole occur at the March and September equinoxes, the station is typically closed owing to the excessive cold from the middle of February until the last week in October, resulting in approximately three months of time during which the telescope can neither be accessed, nor used for CMB observations.

To extend the effective observing season, a Sun shield was constructed during the austral summer of 2001–02, and was in place throughout the 2002 observations. The lightweight conical shield attaches rigidly to the front face of the telescope, completely enclosing, but not rotating with, the faceplate. With a diameter at the attachment point of 62.9 inch (1 inch = 2.54×10^{-2} m), and an opening angle of 15° , the shield prevents any sidelobe of an antenna from seeing the Sun when the telescope is pointed $>65^\circ$ away.

Contamination from the Sun was plainly evident in the 2000 data as fringing in the calibrator visibility data taken after sunrise, and is evident as increased scatter on the shortest baselines even in the 2001 data taken with the ground shield present. These data show excursions in the visibility amplitude on the shortest baselines of $\sim 50\%$ about the pre-sunrise means, although the longest baselines are unaffected. In data taken during the 2002 sunset with the Sun shield present, there is still some evidence for increased scatter in the short-baseline visibilities, but the effect is reduced by nearly two orders of magnitude. For fields $>90^\circ$ from the Sun, the visibility scatter is consistent with the noise on all baseline lengths.

The Sun shield was left in place throughout the 2002 winter season, and it is found that during periods when the Moon is above the horizon but below 10° elevation, the data from 2002 show evidence of excess noise, whereas the 2001 data do not. We speculate that during periods when the Moon is within the opening angle of the Sun shield, secondary reflection off the shield can actually contaminate the short-baseline data when the Moon would otherwise be screened by the ground shield. Accordingly, a stricter cut on the Moon elevation is applied to the 2002 data than to the 2001 data (see the subsection ‘Calibrated visibility cuts’, below).

Polarization response

To derive the polarization response of an interferometer, consider the complex electric field incident on a single receiver, E_m . The response of the interferometer on a baseline $m-n$, called the ‘visibility’, is simply the time-averaged cross-correlation of the fields from a pair of receivers (m, n):

$$V_{mn} = \langle E_m E_n^* \rangle \quad (1)$$

where $\star$ denotes complex conjugation. Rewriting E_m in terms of its components in an arbitrary coordinate system rotated by an angle ψ relative to a coordinate system fixed to the receivers (that is,

parallactic angle), the usual Stokes parameters can be substituted to obtain

$$\begin{aligned} V_{mn}^{RR} &= \frac{1}{2} \{S_I + S_V\} \\ V_{mn}^{LL} &= \frac{1}{2} \{S_I - S_V\} \\ V_{mn}^{RL} &= \frac{1}{2} \{S_Q + iS_U\} e^{-2i\psi} \\ V_{mn}^{LR} &= \frac{1}{2} \{S_Q - iS_U\} e^{2i\psi} \end{aligned} \quad (2)$$

for receivers set to receive right circular (R) and left circular (L) polarization, where S_x is the flux in Stokes parameter x (for $x = I, V, Q$ or U). We refer to these four combinations as ‘Stokes states’, by analogy with the Stokes parameters that can be derived from them.

Because the DASI correlator was originally designed to accommodate only a single Stokes state per receiver, observation in all four states is achieved via time-multiplexing. The polarizer for each receiver is switched between L and R on a sequence specified by a unique Walsh function. This is a periodic two-state function that spends an equal amount of time in each state, that is $\sum_i f_i = 0$ if we represent the two states as $f \in \pm 1$. The product of two Walsh functions is another Walsh function, whence $\sum_{i,n \neq m} f_i^n f_i^m = 0$, and the baseline therefore spends an equal amount of time in the co-polar (RR, LL) and cross-polar (LR, RL) states. In fact, it can be shown that with this sequencing of the polarizers, over the course of a full Walsh cycle each baseline of the interferometer spends an equal amount of time in each of the four Stokes states, although different baselines will sample different Stokes states at any given time. For the observations presented here, a Walsh function of period 16 with a time step of 200 s was used, so that over the course of an hour, every Stokes state is sampled by every baseline for approximately 13 min.

Relative gain and phase calibration

In practice, each receiver introduces a unique phase offset and amplitude variation to the electric field, so that the expression for the detected field must be modified by a receiver-dependent complex gain, $\tilde{E}_m = g_m E_m$, where g_m depends on the polarization state of the receiver. Furthermore, the correlator can introduce baseline-dependent complex gains g_{mn} independent of polarization state, so that the measured visibilities are given by:

$$\begin{aligned} \tilde{V}_{mn}^{RR} &= g_m^R g_n^{R*} g_{mn} V_{mn}^{RR} \equiv G_{mn}^{RR} V_{mn}^{RR} \\ \tilde{V}_{mn}^{LL} &= g_m^L g_n^{L*} g_{mn} V_{mn}^{LL} \equiv G_{mn}^{LL} V_{mn}^{LL} \\ \tilde{V}_{mn}^{RL} &= g_m^R g_n^{L*} g_{mn} V_{mn}^{RL} \equiv G_{mn}^{RL} V_{mn}^{RL} \\ \tilde{V}_{mn}^{LR} &= g_m^L g_n^{R*} g_{mn} V_{mn}^{LR} \equiv G_{mn}^{LR} V_{mn}^{LR} \end{aligned} \quad (3)$$

As the co-polar visibilities $\tilde{V}_{mn}^{RR}$ and $\tilde{V}_{mn}^{LL}$ are just proportional to Stokes $S_I \pm S_V$, the gain factors G_{mn}^{RR} and G_{mn}^{LL} can simply be derived on a per-baseline basis from observations of a bright unpolarized point source, or a more complicated source whose intrinsic visibility structure is known (if we take the circular polarization S_V to be zero).

However, we require a source of known linear polarization to derive the complex gains in the same manner for the cross-polar visibilities $\tilde{V}_{mn}^{RL}$ and $\tilde{V}_{mn}^{LR}$. These are in general scarce at high frequencies, and furthermore few are well studied in the Southern Hemisphere. The low gain of DASI’s 20-cm feed horns, moreover, makes it impracticable to observe all but the brightest sources, in total intensity or otherwise, and the brightest sources at 30 GHz are typically compact H II regions, whose polarization fractions are expected to be negligible.

As can be seen from equation (3), however, we can in principle derive the cross-polar gains from co-polar observations of an unpolarized source, if we work in ratios of complex gains:

$$\frac{\tilde{V}_{mn}^{LL}}{\tilde{V}_{mn}^{RR}} = \left(\frac{g_m^L}{g_m^R}\right) \left(\frac{g_n^L}{g_n^R}\right)^* \equiv r_m r_n^* \quad (4)$$

in which the correlator gains g_{mn} exactly cancel. For DASI, this yields an over-constrained set of 78 complex equations that can be solved for the 13 antenna-based ratios $r_k = g_k^L/g_k^R$.

The solution is not unique, and we must arbitrarily specify the phase $\Delta\phi \equiv \phi_R - \phi_L$ of one of the ratios, typically setting it to zero. With the ratios r_k in hand, the baseline complex gains for the cross-polarized channels can be recovered from the co-polar solutions as:

$$r_n^* G_{mn}^{RR} = \left(\frac{g_m^L}{g_m^R}\right)^* g_m^R g_n^{R*} g_{mn} = e^{-i\Delta\phi} G_{mn}^{RL}$$

$$r_m G_{mn}^{RR} = \left(\frac{g_m^L}{g_m^R}\right) g_m^R g_n^{R*} g_{mn} = e^{i\Delta\phi} G_{mn}^{LR} \quad (5)$$

that is, we can construct the cross-mode calibration factors from the antenna-based solutions, but only up to the unknown phase difference between R and L for our reference antenna. This overall phase offset, while leaving the amplitude of the cross-polar visibilities unaffected, will mix power between S_Q and S_U , or between E and B , in terms of the decomposition into the coordinate-independent scalar fields which has become the standard in the CMB polarization literature^{6–8}. Removal of the offset is therefore critical to obtaining a clean separation of CMB power into E - and B -modes.

In practice, we use the bright, compact, unpolarized H II region, RCW38, to determine the co-polar and cross-polar calibration factors in this manner. The phase offset can only be determined by observing a source whose plane of polarization is known, as described in the next subsection.

Absolute phase calibration

To determine the phase offset introduced by the complex gain correction described above, we require a source whose polarization angle (though not amplitude) is known. As discussed in the previous subsection, however, few suitably strong polarized

sources are available, and in practice we create a source of known polarization angle by observing an unpolarized source through polarizing wire grids. At centimetre wavelengths, the required wire spacing ($\lesssim \lambda/3$) poses no particular technical challenges, and grids can be simply constructed out of conventional materials; 13 wire grids were constructed out of 34-gauge copper wire, wound with a spacing of 0.05 inch. These grids attach directly to the exterior of the corrugated shrouds and completely cover the aperture of the DASI horns.

The same source used for gain calibration of the visibilities, RCW38, was observed with the wire grids attached in two orientations, one with wires parallel to the ground, and the other with grids (but not receivers) rotated by 45° with respect to the first. It was confirmed that in the second orientation the apparent phase of the cross-polar visibilities shifted by $\pm 90^\circ$ relative to the first, as indicated by equation (2).

The measured absolute phase offsets $\Delta\phi(\nu)$ are shown in Fig. 3. These offsets show a trend in frequency ν which is a direct consequence of the achromatic polarizer design discussed in the subsection ‘Receiver polarizer hardware’, above. As Fig. 3 shows, this trend is in excellent agreement with expectations from benchtop measurements of the polarizers, where the expected trend is shown with an arbitrary offset subtracted. These phase offsets were measured in August 2001 and again in February 2002, and were found to agree within measurement errors. As discussed below, we can restrict the intrinsic polarization fraction of RCW38 to $<0.09\%$ at our observing frequency; as the grids create a polarized source with an effective polarization fraction of 100%, any intrinsic polarized flux at this level will have a negligible effect on the measurements of the absolute phase offsets.

Although suitably bright point sources of known polarization angle are not available, we can in principle derive the absolute phase offsets from observations of the Moon, as discussed in the subsection ‘Observations of the Moon’, below. These results, also shown in Fig. 3, are consistent with the offsets measured with the wire grid polarizers, to within the $\sim 0.4^\circ$ measurement errors. As discussed in ref. 4, errors in the absolute phase offset of this magnitude have a negligible effect on the power spectral analysis.

Instrumental polarization

Circular polarization can be thought of as the superposition of two orthogonal linear modes, $E = \alpha E_x + \beta E_y$, where one mode leads the other by exactly 90° , that is, $\beta/\alpha = e^{\pm i\pi/2}$. For perfect polarizers, the cross-polar visibilities are strictly linear combinations of the linear polarization measures, Stokes S_Q and S_U , as indicated by equation (2). For realistic polarizers, however, a small amount of radiation of one handedness will be passed by a polarizer set to transmit the opposite handedness, leading to a term in the expression for the cross-polar visibilities proportional to the total intensity. In the limit of small deviations from perfect orthogonality, $\beta/\alpha \approx i(1 - \epsilon)$, and keeping only terms to first order in ϵ , it can be shown that the cross-polar visibilities are given by:

$$V_{mn}^{RL} = \frac{1}{2} \{ (S_Q + iS_U) e^{-2i\psi} + S_I (\epsilon_m^R + \epsilon_n^{L*}) \}$$

$$V_{mn}^{LR} = \frac{1}{2} \{ (S_Q - iS_U) e^{2i\psi} + S_I (\epsilon_m^L + \epsilon_n^{R*}) \} \quad (6)$$

The terms $L^{RL} \equiv \epsilon_m^R + \epsilon_n^{L*}$ and $L^{LR} \equiv \epsilon_m^L + \epsilon_n^{R*}$ are sums of the polarizer leakages discussed in the subsection ‘Receiver polarization hardware’, above, and are expected from benchtop measurements of the DASI polarizers to be on average $<1\%$ (see Fig. 1). As is clear from equation (6), the presence of leakages has the effect of mixing S_I into S_Q and S_U , or equivalently, the leakages will mix CMB power from T into E and B (see also ref. 4).

On-axis leakage. As discussed in the section ‘Description of the instrument’ above, the receivers are mounted on a rigid faceplate

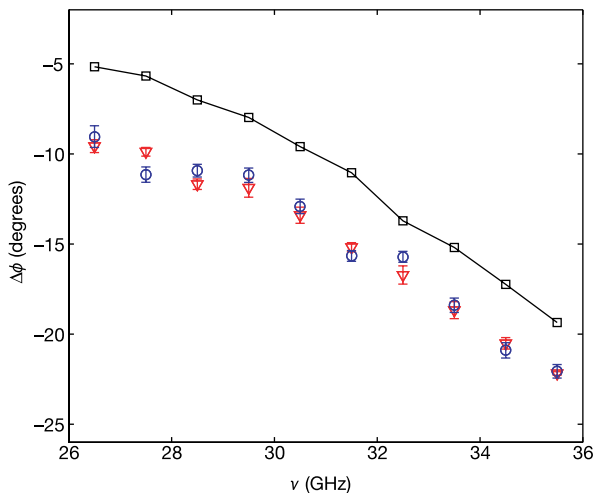


Figure 3 DASI absolute phase offsets. Shown are phase offsets determined from wire grid observations (triangles) and from the Moon (circles). Also shown are benchtop measurements of the same polarizer (solid line, squares), with an arbitrary offset subtracted.

that can be driven to simulate parallactic angle rotation. From equation (6), it can be seen that the leakages are irrotational, while any contribution to the visibilities from intrinsic source polarization will have a phase modulated by the faceplate rotation. The data can therefore be fitted for an offset plus a sinusoidal modulation to isolate the leakages. Equivalently, when observations are made at three- or six-fold symmetric faceplate rotations, the cross-polar visibilities for a radially symmetric source can simply be averaged over faceplate rotations to cancel the intrinsic term.

To determine the magnitude of the leakages, RCW38 was observed at six faceplate rotations separated by 60° , for approximately 3 h at each faceplate position, resulting in an r.m.s. noise on the baseline leakages of 0.8% (of S_I). As can be seen from equation (6), however, the baseline leakages are simply linear combinations of antenna-based terms, and can therefore be solved for the antenna-based leakages ϵ , yielding a typical error of 0.3% on the antenna-based leakages. Typical antenna-based leakage amplitudes range from $\leq 1\%$ over much of the frequency band, to $\sim 2\%$ at the highest frequency. The averages of the antenna-based leakages $\epsilon^{R,L}$ are shown in Fig. 1. Before installation on the telescope, the polarizers were optimized in a test set-up to minimize the leakage response; as can be seen in Fig. 1, these observations indicate that we have achieved close to the theoretical minimum. As discussed in ref. 4, leakages of this magnitude have a negligible effect on the separation of the CMB signal into E and B modes.

Of critical importance is the stability of the leakages, as they cannot be frequently measured. The leakages were measured once in August 2001, and again in April 2002 and July 2002, and the baseline leakages show excellent agreement between all three epochs. The higher signal/noise (s/n) antenna-based leakages show similarly good agreement between August 2001 and April 2002, with residuals at all frequencies $\leq 1\%$, with the exception of three receivers. The polarizers for these receivers were retuned during the 2001–02 austral summer, and systematic offsets can clearly be seen in the antenna-based residuals between these two epochs, the largest being approximately 2.5% in amplitude. Residuals between the April 2002 and July 2002 leakages are $\leq 1\%$ at all frequencies. Variations in the leakages of this magnitude are expected to have a negligible impact on the analysis presented in ref. 4.

As discussed above, the parallactic angle modulation can be used to eliminate the leakage and place a limit on the intrinsic source polarization. Averaging over the three epochs at which the leakages were measured, it is found that the polarization amplitude of RCW38, $P = (S_Q^2 + S_U^2)^{1/2}$, is less than 0.09% (of S_I) at all frequencies.

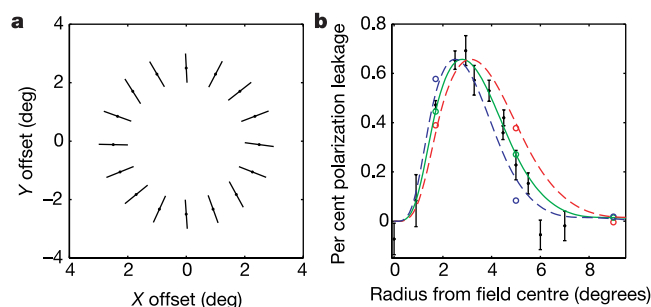


Figure 4 DASI off-axis leakage measurements. **a**, The measured direction of the instrumental leakage for a ring of 2.5° offset pointings on RCW38. **b**, The amplitude profile of the instrumental polarization versus radius from field centre. Moon data are shown as red, green and blue circles corresponding to the frequency bands 26–29, 29–33 and 33–36 GHz, respectively. RCW38 data for the full 26–36 GHz band are shown as black points. The green line is a spline interpolation of the 29–33 GHz Moon data, and the red and blue dashed lines are the interpolation scaled to 27.5 and 34.5 GHz, respectively. See text for further discussion.

Off-axis leakage. Although the polarizers were optimized to yield low leakage, the lensed feed horns themselves will induce an instrumental polarization that varies across the primary beam. The instrumental polarization of the circularly symmetric lenses and feeds is expected to be azimuthally symmetric and depend only on the angle from the phase centre.

To probe the off-axis response of the feeds, observations of RCW38 were made with the source at 16 offset positions in a ring of 2.5° radius. The direction of the instrumental leakage is shown in Fig. 4a. Within the measurement uncertainty the data are consistent with a simple radial pattern. In terms of S_Q and S_U , a radial pattern produces a quadrupole asymmetry across the primary beam, with the S_Q and S_U patterns rotated by 45° with respect to one another.

The radial amplitude profile of this off-axis leakage was characterized by observations of RCW38 and the Moon at various offsets (see Fig. 4). The Moon data have extremely high s/n , permitting a separation of the data into several frequency bands. Attempts to model the off-axis leakage data as a simple perturbation of the aperture field reproduce the general shape and frequency dependence of the measured profile, but fail to replicate the sharp rise in amplitude seen between 1° and 2° . We therefore fit an empirical model by spline interpolation of the Moon data for the 29–33 GHz band (together with a point from RCW38 at 0.9° offset). The effect is assumed to scale with frequency in the same manner as the aperture field, and indeed this is supported by the Moon data at other frequencies. Additionally, the beam-offset RCW38 data confirms the validity of the model, agreeing within the measurement uncertainty.

The presence of off-axis leakage will transform CMB power from S_I into S_Q and S_U , and will bias the estimation of polarization parameters from the data. These effects were investigated via extensive Monte Carlo simulation using the off-axis leakage model shown in Fig. 4. As is shown in ref. 4, the resulting bias on the various likelihood results presented in that paper is $\leq 4\%$ if the off-axis leakage is not accounted for in the analysis. Including our model of the off-axis leakage in the analysis, we can eliminate the bias to 1% or better.

Off-axis leakage will also transform power from point sources in the anisotropy fields into a polarization signal. In ref. 4, we discuss the results of extensive simulations of this effect using the best available point source counts at centimetre wavelengths. The results indicate that the bias on the derived CMB parameters is again small, of order a few per cent or less.

Polarization observations with DASI

During the course of the 2001–02 CMB observations, considerable effort was expended to characterize the polarization response of the instrument through observations of astronomical sources. In the preceding sections, we have described at length observations of the unpolarized H II region RCW38, on which the bulk of DASI's calibration is based. Here we present polarimetric observations of a variety of other sources that provide independent checks of those calibrations, and lend confidence that we understand the various aspects of the instrumental response to a level (precision better than 1%) beyond that required to measure polarization in the CMB.

Observation of the Moon

At centimetre wavelengths, radiation from the Moon is dominated by thermal emission from the regolith, typically from the first ~ 10 – 20 cm of the Moon's surface. This radiation is intrinsically unpolarized, but scattering off the dielectric discontinuity at the surface will induce polarization; the tangentially polarized component is preferentially reflected, leading to a net radial polarization pattern across the disk of the Moon. This polarization amplitude decreases to zero at the centre and increases to a maximum at the limb (see ref. 9 for a comparable result at 21 cm, ref. 10 for a similar discussion

and map of the microwave polarization of Mercury, and ref. 11 for Moon measurements at 8 GHz).

Observations of the Moon were made with DASI during 2001–02 at several epochs, and at various phases of the Moon. All show excellent agreement with the expected radial polarization pattern, and the consistency between epochs attests to the stability of the absolute phase offset. In Fig. 5 we present a polarized map of the Moon made during 21–22 August 2001, when the Moon was nearly full. These data have been corrected for instrumental leakages and the absolute phase offset determined from wire grid observations, discussed in the subsection ‘Absolute phase calibration’, above.

Note that any residual phase offset in the cross-polar visibilities will introduce a vorticity to the polarization vectors, from which we can derive an independent measure of the cross-polar phase offsets. These measurements are shown in Fig. 3, and are in excellent agreement with the phase offsets determined from the wire grid observations discussed in the subsection ‘Absolute phase calibration’ above.

Observations of NGC6334

Polarimetric observations of the Galactic source NGC6334 were made in August 2001, centred on right ascension (RA) 17 h 20 min, declination (dec.) $-35^{\circ} 50'$ (J2000). NGC6334 is a massive molecular ridge, with numerous embedded photo-dissociation regions, exhibiting the rich phenomenology typical of massive star-forming complexes. At radio and infrared wavelengths, emission from this complex encompasses a broad variety of phases of the interstellar medium, ranging from line and continuum emission from molecular gas and dust, to optically thick and dust-obscured free-free emission from ionized gas in embedded H II regions (see, for example, refs 12 and 13).

In Fig. 6, we present maps of the source in total intensity and polarization fraction. Artefacts of the Fourier sampling of the array have been reduced in these maps by an application of the CLEAN algorithm¹⁴. At DASI’s $\sim 21'$ resolution, the microwave emission from the ridge is concentrated in two regions of comparable flux, one on-centre, the second near the half-power point of the primary beam. This source therefore provides an ideal testbed for the leakage corrections described in the subsections ‘On-axis leakage’ and

‘Off-axis leakage’ above. As can be seen in Fig. 6b, the S_Q and S_U maps show emission coincident with these regions, at a level consistent with the instrumental leakages. Application of the on-axis leakage correction to the visibilities completely removes the central source in the polarization map, demonstrating that we can correct for instrumental leakage to better than 0.2%. Applying the off-axis leakage profile in the image plane similarly accounts for the second source.

The residuals after correction are consistent with the noise in the polarization map, indicating that the net polarization fraction from the variety of emission processes represented in the source is $< 0.14\%$.

CMB observations

Field selection

For the total intensity anisotropy measurements presented in ref. 1, we observed four rows of eight fields separated by 1 h in RA. The locations of these fields were chosen as a compromise between observing at high elevation to minimize ground signal, and in regions showing very low levels of Galactic emission in both the IRAS 100 μm and Haslam 408 MHz maps¹⁵. We were able clearly to identify point sources from our own data within these fields as described in ref. 1.

For the polarization observations presented here, we selected the C2 and C3 fields centred at RA 23 h 30 min and 00 h 30 min, and dec. -55° , within which no point sources had been detected (see Fig. 7). These fields lie at Galactic latitude -58.4° and -61.9° , respectively. The brightness of the IRAS 100 μm and 408 MHz maps within our fields lie at the 6% and 25% points, respectively, of the integral distributions taken over the whole sky.

Observations

The fields were tracked over the full azimuth range in 2-h blocks. In any block, each field was tracked for approximately 1 h over the same azimuth range, allowing a constraint on any residual ground signal not removed by the shields. Observations were divided into self-contained 24-h segments. Each segment comprised 20 h of CMB observation, and bracketing observations of the primary calibrator source RCW38. This source has been described at length in ref. 1, and the reader is referred to that paper for details.

During 2000, observations of RCW38 performed every 12 h demonstrated that instrumental gains were stable at the $\sim 1\%$ level over many days; during 2001–02, the source was therefore observed only at the beginning and end of the 20-h CMB observations. In each calibrator scan, the source was observed for 35 min in each of the RR and LL configurations, from which the instrumental gains were determined for both the co-polar and cross-polar visibilities, as discussed in the subsection ‘Relative gain and phase calibration’, above. In 35 min, we achieve an accuracy of $\lesssim 2\%$ on both the co-polar and cross-polar visibilities. These observations were bracketed by skydips from which the opacity was determined, and by injection of a noise source used to calibrate the complex correlator, as described in ref. 1.

To provide an additional level of consistency checks, observations were performed at two distinct rotations of the array faceplate (see the section ‘Description of the instrument’, above), separated by 60° . Each 24-h segment was observed at a fixed faceplate position, but the faceplate position was alternated roughly every day, so that approximately half the data from 2001–02 were acquired in each orientation.

Fields were observed from 10 April 2001 to 27 October 2001, and again from 14 February 2002 to 11 July 2002. In all, 271 days of usable CMB data were acquired, 162 from 2001, and 109 from 2002, with the bulk of the remaining time spent on the calibration observations described above, as well as periodic observations to determine the pointing model of the telescope. As discussed in the subsection ‘On-axis leakage’ above, observations to determine the

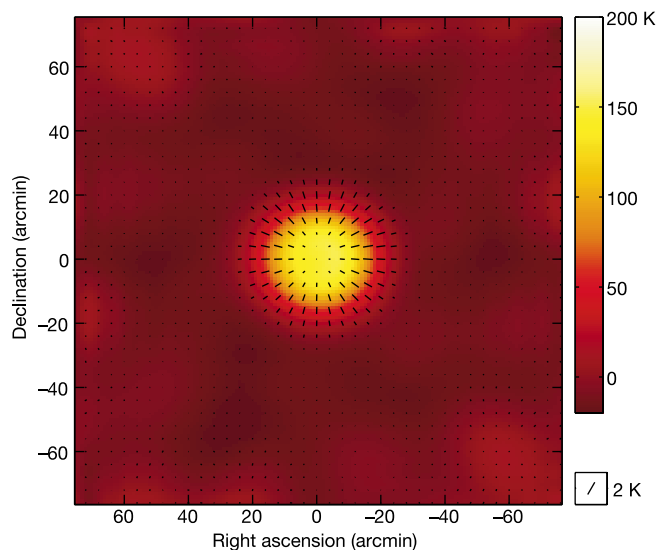


Figure 5 DASI image of the Moon. Shown is the total intensity map (colour scale), with measured polarization vectors overplotted, demonstrating the expected radial polarization pattern. This is a direct Fourier transform of the visibility data. Conversion to temperature units assumes a $\sim 21'$ (FWHM) gaussian for the main lobe of the synthesized beam and is therefore only approximate.

on-axis leakages were repeated at three epochs to check the stability of the instrumental polarization. Observations to determine the cross-polar phase offset discussed in the subsection ‘Absolute phase calibration’, above, were repeated twice to check for stability. Likewise observations of the Moon were repeated on three separate occasions during this time period.

Data reduction

As described in ref. 1, raw visibility data from the correlators are accumulated in 8.4-s integrations, along with monitoring data from various hardware systems throughout the telescope. After editing and calibration, the raw visibilities for each 24-h period are combined, and these daily averages are the inputs to the analysis presented in ref. 4. ‘Edits’ applied to the data can be broadly divided into hardware and calibration edits applied to the raw integrations, and edits applied to the 24-h averages, and we describe each of these in turn.

Calibration and raw visibility cuts. As the first level of editing, the 8.4-s integrations are rejected during periods when a receiver local oscillator was out of lock, when a receiver was warm, or when the total power for a given receiver was outside the normal range. These cuts collectively reject $\sim 11\%$ of the data.

The only cut based on the visibility data values corresponds to a $>30\sigma$ outlier rejection, to remove rare hardware glitches; this cut rejects $\ll 0.1\%$ of the data.

The surviving 8.4-s integrations from each 24-h period are then calibrated for the relative amplitude and phase of the complex multipliers on the correlator cards, determined by the periodic injection of a broad-band noise source. These calibration factors have proved to be extremely repeatable in the three years that the telescope has operated, but we nevertheless reject baselines for which these factors lie on the tails of their distributions, indicating possible problems with the multiplier hardware. Data are rejected when the relative gain of the real and imaginary multipliers falls outside the range 0.6–1.2, or when the relative phase exceeds $\pm 20^\circ$, regimes for which the distributions of these factors over all baselines have become noticeably non-gaussian. These cuts reject an additional 13% of the data.

As described in the subsection ‘Polarization response’ above, every baseline samples the four Stokes states on a Walsh sequence. The observations presented here employ a period-16 Walsh sequence with a time step of 200 s, so that an observation is nominally completed in 3,200 s. With the inclusion of switching time for the polarizers, however, the net observing time is somewhat less, and a reduction of >200 s is generally an indication of a polarizer sticking (see the discussion of the polarizer assembly in the subsection ‘Receiver polarization hardware’, above). In combination with the cuts above, we therefore reject any scan with total time in a valid Stokes state $<3,000$ s, rejecting 0.1% of the data. Calibrator scans are also rejected if previous edits have reduced the total

integration time by more than a factor of two. Loss of a calibrator scan results in the rejection of an entire day of data.

The data are next calibrated to remove complex instrumental gains, and to convert to absolute flux scale, using the bracketing observations of RCW38. This gain calibration is equivalent to the combination of the amplitude and phase calibration described in ref. 1, but here we construct the cross-polar gains from antenna-based fits to the co-polar visibilities, as described in the subsection ‘Relative gain and phase calibration’, above. The flux scale of RCW38 has been previously determined to 3.5% from measurements of absolute loads transferred to the source in 2000 and 2001, as discussed in ref. 1.

All data that lack bracketing calibrator scans are rejected, as are co-polar data for which the calibrator amplitude varies by $>10\%$ over 24 h, or for which the calibrator phase varies by $>15^\circ$. For the cross-polar data, these limits are relaxed to 15% and 20° , as the lower s/n cross-polar data are dominated by thermal noise, and are therefore relatively insensitive to random errors in the daily calibration. These cuts reject 5% of the data.

Calibrated visibility cuts. After the raw data are edited and calibrated, the 8.4-s integrations are combined to form 24-h averages of the visibilities for every baseline. (Recall that for a coplanar array, the projected baseline lengths do not change as a source is tracked across the sky, and that sources at the South Pole do not rotate in parallactic angle. The combination of visibilities on arbitrary timescales therefore engenders no loss of information due to smearing in the Fourier plane.) Cuts applied to the 24-h averages include astronomical cuts based on the position of the Sun and Moon, weather cuts, and cuts derived from the visibility noise.

In the first category, co-polar visibilities are cut whenever the Sun was above the horizon. The cross-polar visibilities show little evidence for contamination from the Sun at low elevations, and are rejected only when the Sun elevation exceeds 5° , or the Sun is closer than 90° to the fields and above the horizon.

The Moon rises and sets once a month at the South Pole, and passes within 45° of the CMB fields. Although the data show little evidence of contamination for Moon elevations $<10^\circ$, co-polar data were rejected when the Moon was above the horizon and closer than 80° and 60° , respectively, to the CMB and calibrator positions. A less stringent cut is imposed on the cross-polar data, as they are dominated by thermal noise and are therefore less sensitive to daily calibration uncertainties. For the 2002 data, there is evidence that the Sun shield may actually increase the data scatter through secondary reflections when the Moon would otherwise be below the ground shield, and the 2002 co-polar data are rejected whenever the Moon is above the horizon.

The correlation matrix of visibilities was computed for each day of data, over all 78 complex baselines, 10 correlator frequencies, and 4 Stokes states, where concurrent data from different Stokes states are available (see the subsection ‘Polarization response’, above). This

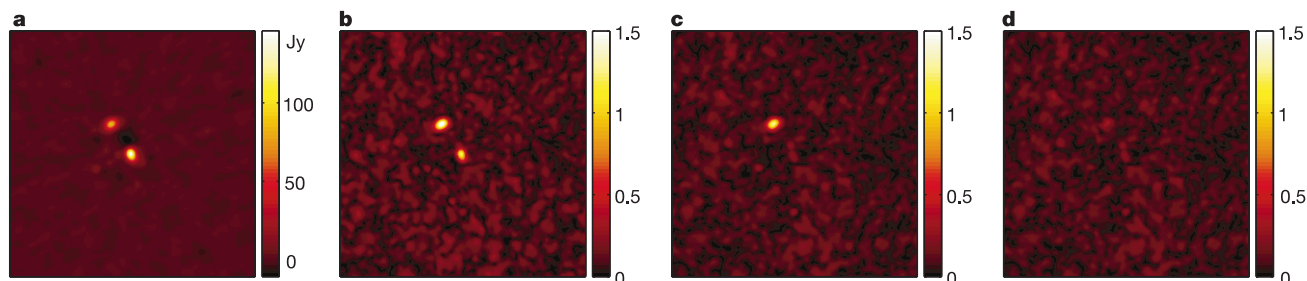


Figure 6 Observations of the molecular cloud complex NGC6334. **a**, Total intensity map. **b**, Map of $P = (S_0^2 + S_V^2)^{1/2}$, uncorrected for leakage, showing structure correlated with the total intensity map, and at a level consistent with instrumental polarization. **c**, Map of P corrected for on-axis leakage. The central source has disappeared, and the on-axis

residuals are consistent with intrinsic source polarization $<0.14\%$. **d**, Map of P , corrected for off-axis leakage, using the profiles shown in Fig. 4. Fields are 12.8° across, with RA increasing to the left. All intensity scales are in Jy.

$6,240 \times 6,240$ matrix is inspected for off-diagonal correlations, and we reject entire days when the significance of any correlation exceeds 8σ , presumably due to weather. This cut rejects 22 days of data not already rejected by the lunar and solar cuts (8.1% of the data).

As an additional test of the correlator hardware, we form the sum of visibilities over consecutive pairs of 1-h observations, averaged over 24 h, and over all baselines for each of DASI's 10 correlators. (Note, each correlator performs the 78 real and imaginary correlations for one of the ten 1-GHz-wide DASI frequency bands, see refs 16 and 1.) This quantity is sensitive to output of the correlator that is coherent between observations of the two CMB fields. All data are rejected for an entire correlator on days when there is evidence for offsets large compared to the thermal noise. Not surprisingly, this statistic rejects data for all correlators taken during sunset 2002, but also rejects data from a single correlator that showed unusually large offsets throughout much of 2002. Of the data not previously rejected from the astronomical and weather cuts described above, this edit rejects an additional 1.9%.

Data are also rejected from an entire correlator on days when the 1-h variance, averaged over all baselines of that correlator, is grossly discrepant with the mean variance over 2001–02. Data are rejected on a per-baseline basis if the variance within any 1-h period falls on the extreme tails of the expected reduced χ^2 distribution. Collectively, these edits reject a negligible fraction of the data.

For all of the cuts described above, the results are insensitive to small variations in the chosen cut levels. As discussed in ref. 4, as these cut levels are varied, evidence for residual contamination eventually appears first in χ^2 consistency tests on differenced data sets, then as excess noise in the 1-h observations, and only lastly in the likelihood results. The likelihood results are found to be insensitive even to quite large variations in the precise edit thresholds, and we are confident that the threshold settings introduce no significant bias in the CMB analysis.

CMB maps and discussion

In this experiment, two CMB fields were observed to provide a constraint on any common signal not removed by the ground shield; accordingly, the analyses presented in ref. 4 pertain to the differenced visibilities of the two CMB fields. As described in that paper, the differenced visibilities are subjected to numerous consistency tests and no evidence is found for excess noise in any of the various subdivisions of the data by epoch, deck rotation or as a function of baseline length. Here we present maps of the differenced co-polar visibility data; discussion of the cross-polar data is presented in ref. 4.

Shown in Fig. 8 are total intensity maps of the differenced visibilities, where the maps are the sum and difference of two epochs with equal amounts of data. These maps have been re-normalized so that the variance in the maps is appropriate for a

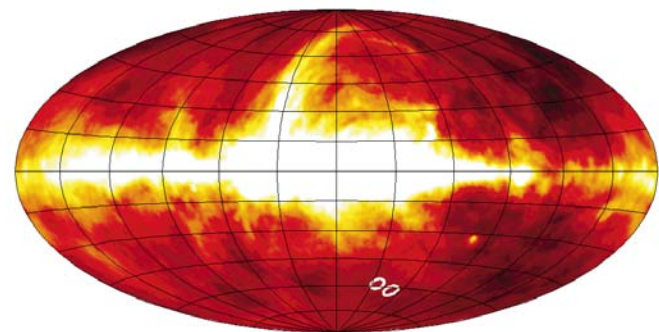


Figure 7 Polarization field locations. DASI fields are marked by white circles, plotted over the 408 MHz synchrotron map¹⁵. The colour scale is linear, and has been clipped at one-tenth of the maximum to allow the structure at high Galactic latitudes to be seen.

single field. These maps are the direct Fourier transform of all the visibility data, with natural weighting. They are therefore directly comparable to maps made by chopped or swept-beam total power techniques, and represent the actual sky structure convolved with the synthesized beam pattern. The epoch-sum map clearly shows structure enveloped by the primary beam of the feed horns, and as expected from the consistency tests, no residual signal is seen in the epoch-difference map above the noise. The detected structure is also consistent with visibilities measured for these fields in the 2000 data presented in ref. 1. With a noise floor in the difference map of $2.8 \mu\text{K}$, we achieve a s/n ratio of approximately 25 at beam centre, for a synthesized beam of $\sim 21'$ (full-width at half-maximum, FWHM).

In ref. 1, we gave estimates of the total intensity of synchrotron, free-free and thermal dust emission in the region of our fields, showing that the expected amplitudes are very small. This was confirmed in ref. 2 by a template-based cross-correlation analysis which demonstrated that the contribution of each of these foregrounds to our temperature anisotropy results was negligible.

No point sources are detected in either field, and there is at best scant evidence for enhanced emission at the locations of the 31 sources in the Parkes-MIT-NRAO (PMN) catalogue¹⁷ that are most likely to contribute. Nevertheless these sources, whose 5 GHz flux when enveloped with the DASI primary beam exceeds 50 mJy, are projected out of the co-polar data for all of the likelihood results presented in ref. 4.

As discussed in ref. 4, with these sources projected out of the differenced co-polar data, the temperature spectral index is consistent with a thermal spectrum, $\beta = -0.01$ (-0.16 to 0.14 at 68% confidence), where $T \propto \nu^\beta$. Any significant non-thermal emission

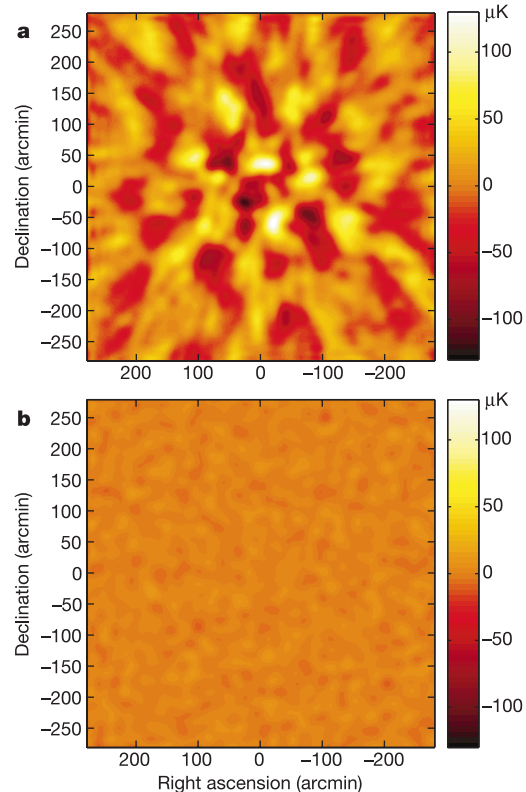


Figure 8 Total intensity maps of the differenced (C2–C3) CMB fields. Differenced visibilities are given by $\Delta = (V_{C2} - V_{C3})/\sqrt{2}$. **a**, Epoch-sum map $(\Delta_1 + \Delta_2)/2$ of two epochs containing equal amounts of data. **b**, Epoch-difference map $(\Delta_1 - \Delta_2)/2$ of the two epochs, plotted to the same temperature scale. The noise level in the map is $2.8 \mu\text{K}$, for a s/n at beam centre of ~ 25 . These maps are made by direct Fourier transform of all of the visibility data, with natural weighting.

can therefore be excluded with high confidence. If none of the PMN sources are projected out, the spectral index limit shifts to -0.12 (-0.25 to 0.01), indicating that the PMN sources may be detected statistically, but that the CMB is nevertheless the dominant signal in these fields.

We conclude that the structure seen in the temperature map is consistent with the CMB and inconsistent with any known source of foreground emission. In 271 days of observation, DASI has obtained a detection of CMB anisotropy with unprecedented sensitivity. Careful optimization of the polarizers and extensive calibration of the instrument have resulted in characterization of the instrumental polarization response to levels well below what is required for a detection of CMB polarization. □

Received 7 October; accepted 5 November 2002; doi:10.1038/nature01271.

- Leitch, E. M. *et al.* Experiment design and first season observations with the Degree Angular Scale Interferometer. *Astrophys. J.* **568**, 28–37 (2002).
- Halverson, N. W. *et al.* Degree Angular Scale Interferometer first results: A measurement of the cosmic microwave background angular power spectrum. *Astrophys. J.* **568**, 38–45 (2002).
- Pryke, C. *et al.* Cosmological parameter extraction from the first season of observations with the Degree Angular Scale Interferometer. *Astrophys. J.* **568**, 46–51 (2002).
- Kovac, J. M. *et al.* Detection of polarization in the cosmic microwave background using DASI. *Nature* **420**, 772–787 (2002).
- Padin, S. *et al.* The Cosmic Background Imager. *Publ. Astron. Soc. Pacif.* **114**, 83–97 (2002).
- Kamionkowski, M., Kosowsky, A. & Stebbins, A. Statistics of cosmic microwave background polarization. *Phys. Rev. D* **55**, 7368–7388 (1997).
- Zaldarriaga, M. & Seljak, U. All-sky analysis of polarization in the microwave background. *Phys. Rev. D* **55**, 1830–1840 (1997).
- Hu, W. & White, M. A CMB polarization primer. *New Astron.* **2**, 323–344 (1997).
- Moffat, P. H. Aperture synthesis polarimetry of the moon at 21 cm. *Mon. Not. R. Astron. Soc.* **160**, 139–154 (1972).
- Mitchell, D. I. & De Pater, I. Microwave imaging of Mercury's thermal emission at wavelengths from 0.3 to 20.5 cm. *Icarus* **110**, 2–32 (1994).
- Poppi, S., Carretti, E., Cortiglioni, S., Krotikov, V. D. & Vinyajkin, E. N. *AIP Conf. Proc.* **609**, *Astrophysical Polarized Backgrounds* (eds Cecchini, S., Cortiglioni, S., Sault, R. & Sbarra, C.) 187–192 (AIP, Melville, New York, 2002).
- Burton, M. G. *et al.* High-resolution imaging of photodissociation regions in NGC 6334. *Astrophys. J.* **542**, 359–366 (2000).
- Jackson, J. M. & Kraemer, K. E. Photodissociation regions and HII regions in NGC 6334. *Astrophys. J.* **512**, 260–270 (1999).
- Högbom, J. A. Aperture synthesis with a non-regular distribution of interferometer baselines. *Astron. Astrophys. Suppl.* **15**, 417–426 (1974).
- Haslam, C. G. T. *et al.* A 408 MHz all-sky continuum survey. I—observations at southern declinations and for the north polar region. *Astron. Astrophys.* **100**, 209–219 (1981).
- Padin, S., Cartwright, J. K., Shepherd, M. C., Yamasaki, J. K. & Holzapfel, W. L. A wideband analog correlator for microwave background observation. *IEEE Trans. Instrum. Meas.* **50**, 1234–1240 (2001).
- Wright, A. E., Griffith, M. R., Burke, B. F. & Ekers, R. D. The Parkes-MIT-NRAO (PMN) surveys. 2: Source catalog for the southern survey ($-87.5^\circ < \delta < -37^\circ$). *Astrophys. J. Suppl.* **91**, 111–308 (1994).

Acknowledgements We are indebted to the Caltech CBI team led by T. Readhead, in particular to S. Padin for a considerable fraction of DASI's electronics design, to J. Cartwright for the downconverter design and to M. Shepherd for data acquisition software. E. Chauvin of Vertex/RSI is thanked for professional excellence in the design and delivery of the telescope mount. We thank Raytheon Polar Services for their support of the DASI project, including the erection of the DASI ground shields, and in particular R. Rowatt and his crew. We are indebted to the Center for Astrophysical Research in Antarctica (CARA), in particular the efforts of A. Day, S. Meyer, N. Odalen, R. D. and E. Pernic, R. Lowenstein, R. Spotz, M. Whitehead and the CARA polar operations staff. We thank J. Kooi for his assistance with the installation of the polarization hardware, J. Yamasaki for help with electronics, E. LaRue and G. Drag for the assembly of the calibration wire grids, M. Loh for assistance with the Sun shield and K. Coble for analysis of DASI primary beam measurements. We thank the TopHat collaboration for the use of their Sun shield. We thank the observatory staff of the Australia Telescope Compact Array, in particular B. Sault and R. Subrahmanyan, for their generosity in providing point source observations of the DASI fields. This research was initially supported by the National Science Foundation (NSF) under a cooperative agreement with CARA, a NSF Science and Technology Center; it is currently supported by the NSF-OPP. J.E.C. acknowledges support from the James S. McDonnell Foundation and the David and Lucile Packard Foundation. J.E.C. and C.P. acknowledge support from the Center for Cosmological Physics.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to E.M.L. (e-mail: eml@polestar.uchicago.edu).

Detection of polarization in the cosmic microwave background using DASI

J. M. Kovac^{*†‡}, E. M. Leitch^{†‡}, C. Pryke^{†‡||}, J. E. Carlstrom^{*†‡||}, N. W. Halverson^{†‡} & W. L. Holzapfel^{†‡}

^{*} Department of Physics; [†] Center for Astrophysical Research in Antarctica; [‡] Center for Cosmological Physics; [§] Department of Astronomy & Astrophysics and ^{||} Enrico Fermi Institute, University of Chicago, 5640 South Ellis Avenue, Chicago, Illinois 60637, USA

[¶] Department of Physics, University of California at Berkeley, Le Conte Hall, California 94720, USA

The past several years have seen the emergence of a standard cosmological model, in which small temperature differences in the cosmic microwave background (CMB) radiation on angular scales of the order of a degree are understood to arise from acoustic oscillations in the hot plasma of the early Universe, arising from primordial density fluctuations. Within the context of this model, recent measurements of the temperature fluctuations have led to profound conclusions about the origin, evolution and composition of the Universe. Using the measured temperature fluctuations, the theoretical framework predicts the level of polarization of the CMB with essentially no free parameters. Therefore, a measurement of the polarization is a critical test of the theory and thus of the validity of the cosmological parameters derived from the CMB measurements. Here we report the detection of polarization of the CMB with the Degree Angular Scale Interferometer (DASI). The polarization is detected with high confidence, and its level and spatial distribution are in excellent agreement with the predictions of the standard theory.

The CMB radiation provides a pristine probe of the Universe roughly 14 billion years ago, when the seeds of the complex structures that characterize the Universe today existed only as small density fluctuations in the primordial plasma. As the physics of such a plasma are well understood, detailed measurements of the CMB can provide critical tests of cosmological models and can determine the values of cosmological parameters with high precision. The CMB has accordingly been the focus of intense experimental and theoretical investigations since its discovery nearly 40 years ago¹. The frequency spectrum of the CMB was well determined by the COBE FIRAS instrument^{2,3}. The initial detection of temperature anisotropy was made on large angular scales by the COBE DMR instrument⁴ and recently there has been considerable progress in measuring the anisotropy on finer angular scales^{5–10}.

In the now standard cosmological model (see, for example, ref. 11), the shape of the CMB angular power spectrum directly traces acoustic oscillations of the photon-baryon fluid in the early Universe. As the Universe expanded, it cooled; after roughly 400,000 years, the intensity of the radiation field was no longer sufficient to keep the Universe ionized, and the CMB photons decoupled from the baryons as the first atoms formed. Acoustic oscillations passing through extrema at this epoch are observed in the CMB angular power spectrum as a harmonic series of peaks. Polarization is also a generic feature of these oscillations^{12–18}, and thus provides a model-independent test of the theoretical framework^{19–21}. In addition, detection of polarization can in principle triple the number of observed CMB quantities, enhancing our ability to constrain cosmological parameters.

CMB polarization arises from Thompson scattering by electrons of a radiation field with a local quadrupole moment²². In the primordial plasma, the local quadrupole moment is suppressed until decoupling, when the photon mean free path begins to grow. At this time, the largest contribution to the local quadrupole is due to Doppler shifts induced by the velocity field of the plasma²³. CMB polarization thus directly probes the dynamics at the epoch of decoupling. For a spatial Fourier mode of the acoustic oscillations, the velocities are perpendicular to the wavefronts, leading only to perpendicular or parallel alignment of the resulting polarization, which we define as positive and negative respectively. We refer to these polarization modes as scalar *E*-modes in analogy with electric fields; they have no curl component. Because the level of the

polarization depends on velocity, we expect that the peaks in the scalar *E*-mode power spectrum correspond to density modes that are at their highest velocities at decoupling and are thus at minimum amplitude. The location of the harmonic peaks in the scalar *E*-mode power spectrum are therefore expected to be out of phase with the peaks in the temperature (*T*) spectrum^{15–17}.

In light of the above discussion, it is clear that the CMB temperature and polarization anisotropy should be correlated at some level²⁴. For a given multipole, the sign of the *TE* correlation should depend on whether the amplitude of the density mode was increasing or decreasing at the time of decoupling. We therefore expect a change in the sign of the *TE* correlation at maxima in the *T* and the *E* power spectra, which correspond to modes at maximum and minimum amplitude, respectively. The *TE* correlation therefore offers a unique and powerful test of the underlying theoretical framework.

Primordial gravity waves will lead to polarization in the CMB^{14,25} with an *E*-mode pattern as for the scalar density perturbations, but will also lead to a curl component, referred to as *B*-mode polarization^{26–28}. The *B*-mode component is due to the intrinsic polarization of the gravity waves. In inflationary models, the level of the *B*-mode polarization from gravity waves is set by the fourth power of the inflationary energy scale. While the detection of *B*-mode polarization would provide a critical test of inflation, the signal is likely to be very weak and may have an amplitude that is effectively unobservable²⁹. Furthermore, distortions to the scalar *E*-mode polarization by the gravitational lensing of the intervening large scale structure in the Universe will lead to a contaminating *B*-mode polarization signal which will severely complicate the extraction of the polarization signature from gravity waves^{30–33}. The possibility, however, of directly probing the Universe at energy scales of order 10¹⁶ GeV by measuring the gravity-wave induced polarization (see, for example, ref. 34) is a compelling goal for CMB polarization observations.

Prior to the results presented in this paper, only upper limits have been placed on the level of CMB polarization. This is due to the low level of the expected signal, demanding sensitive instruments and careful attention to sources of systematic uncertainty (see ref. 35 for a review of CMB polarization measurements).

The first limit to the degree of polarization of the CMB was set in 1965 by Penzias and Wilson, who stated that the new radiation that

they had discovered was isotropic and unpolarized within the limits of their observations¹. Over the next 20 years, dedicated polarimeters were used to set much more stringent upper limits on angular scales of order several degrees and larger^{36–41}. The current best upper limits for the *E*-mode and *B*-mode polarizations on large angular scales are $10\ \mu\text{K}$ at 95% confidence for the multipole range $2 \leq l \leq 20$, set by the POLAR experiment⁴².

On angular scales of the order of one degree, an analysis of data from the Saskatoon experiment⁴³ set the first upper limit ($25\ \mu\text{K}$ at 95% confidence for $l \approx 75$); this limit is also noteworthy in that it was the first limit that was lower than the level of the CMB temperature anisotropy. The current best limit on similar angular scales was set by the PIQUE experiment⁴⁴—a 95% confidence upper limit of $8.4\ \mu\text{K}$ to the *E*-mode signal, assuming no *B*-mode polarization. A preliminary analysis of cosmic background imager (CBI) data⁴⁵ indicates an upper limit similar to the PIQUE result, but on somewhat smaller scales. An attempt was also made to search for the *TE* correlation using the PIQUE polarization and Saskatoon temperature data⁴⁶.

Polarization measurements have also been pursued on much finer angular scales (of the order of an arcminute), resulting in several upper limits (for example, refs 47 and 48). However, at these angular scales, corresponding to multipoles of about 5,000, the level of the primary CMB anisotropy is strongly damped and secondary effects due to interactions with large-scale structure in the Universe are expected to dominate¹¹.

In this paper, we report the detection of polarized anisotropy in the CMB radiation with the Degree Angular Scale Interferometer (DASI) located at the National Science Foundation (NSF) Amundsen–Scott South Pole research station. The polarization data were obtained during the 2001 and 2002 austral winter seasons. DASI was previously used to measure the temperature anisotropy from $140 < l < 900$, during the 2000 season. We presented details of the instrument, the measured power spectrum and the resulting cosmological constraints in a series of three papers: refs 49, 6 and 50. Prior to the start of the 2001 season, DASI was modified to allow polarization measurements in all four Stokes parameters over the same l range as the previous measurements. The modifications to the instrument, observational strategy, data calibration and data reduction are discussed in detail in a companion paper in this issue⁵¹.

The measurements reported here were obtained within two 3.4° full-width at half-maximum (FWHM) fields separated by 1 h in right ascension. The fields were selected from the subset of fields observed with DASI in 2000 in which no point sources were detected, and are located in regions of low Galactic synchrotron and dust emission. The temperature angular power spectrum is found to be consistent with previous measurements and its measured frequency spectral index is -0.01 (-0.16 to 0.14 at 68% confidence), where 0 corresponds to a $2.73\ \text{K}$ Planck spectrum. Polarization of the CMB is detected at high confidence ($\geq 4.9\sigma$) and its power spectrum is consistent with theoretical predictions, based on the interpretation of CMB anisotropy as arising from primordial scalar adiabatic fluctuations. Specifically, assuming a shape for the power spectrum consistent with previous temperature measurements, the level found for the *E*-mode polarization is 0.80 (0.56 to 1.10 68% confidence interval), where the predicted level given previous temperature data is 0.9 to 1.1 . At 95% confidence, an upper limit of 0.59 is set to the level of *B*-mode polarization parameterized with the same shape and normalization as the *E*-mode spectrum. The *TE* correlation of the temperature and *E*-mode polarization is detected at 95% confidence and is also found to be consistent with predictions.

With these results contemporary cosmology has passed a long anticipated and crucial test. If the test had not been passed, the underpinnings of much of what we think we know about the origin and early history of the Universe would have been cast into doubt.

Measuring polarization with DASI

DASI is a compact interferometric array optimized for the measurement of CMB temperature and polarization anisotropy^{49,51}. Because they directly sample Fourier components of the sky, interferometers are well suited to measurements of the CMB angular power spectrum. In addition, an interferometer gathers instantaneous two-dimensional information while inherently rejecting large-scale gradients in atmospheric emission. For observations of CMB polarization, interferometers offer several additional features. They can be constructed with small and stable instrumental polarization. Furthermore, linear combinations of the data can be used to construct quantities with essentially pure *E*- and *B*-mode polarization response patterns on a variety of scales. This property of the data greatly facilitates the analysis and interpretation of the observed polarization in the context of cosmological models.

DASI is designed to exploit these advantages in the course of extremely long integrations on selected fields of sky. The 13 horn/lens antennas that comprise the DASI array are compact, axially symmetric, and sealed from the environment, yielding small and extremely stable instrumental polarization. Additional systematic control comes from multiple levels of phase switching and field differencing designed to remove instrumental offsets. The DASI mount is fully steerable in elevation and azimuth with the additional ability to rotate the entire horn array about the faceplate axis. The flexibility of this mount allows us to track any field visible from the South Pole continuously at constant elevation angle, and to observe it in redundant faceplate orientations which allow sensitive tests for residual systematic effects.

Instrumental response and calibration

Each of DASI's 13 receivers may be set to admit either left or right circular polarization. An interferometer measures the correlations between the signals from pairs of receivers, called visibilities; as indicated by equation (4) in the 'Theory covariance matrix' subsection recovery of the full complement of Stokes parameters requires the correlation of all four pairwise combinations of left and right circular polarization states (RR, LL, RL and LR), which we refer to as Stokes states. The co-polar states (RR, LL) are sensitive to the total intensity, while the cross-polar states (RL, LR) measure linear combinations of the Stokes parameters *Q* and *U*.

Each of DASI's analogue correlator channels can accommodate only a single Stokes state, so measurement of the four combinations is achieved via time-multiplexing. The polarizer for each receiver is switched on a 1-h Walsh cycle, with the result that over the full period of the cycle, every pair of receivers spends an equal amount of time in all four Stokes states.

In ref. 51, we detail the calibration of the polarized visibilities for an interferometer. In order to produce the calibrated visibilities as defined in equation (4) below, a complex gain factor which depends on the Stokes state must be applied to each raw visibility. Although the cross-polar gain factors could easily be determined with observations of a bright polarized source, no suitable sources are available, and we therefore derive the full calibration through observations of an unpolarized source. The gains for a given pair of receivers (a baseline) can be decomposed into antenna-based factors, allowing us to construct the cross-polar gains from the antenna-based gain factors derived from the co-polar visibilities. DASI's calibration is based on daily observations of the bright H II region RCW38, which we described at length in ref. 49. We can determine the individual baseline gains for all Stokes states with statistical uncertainties $< 2\%$ for each daily observation. Systematic gain uncertainties for the complete data set are discussed in the 'Systematic uncertainties' section.

The procedure for deriving the baseline gains from antenna-based terms leaves the cross-polar visibilities multiplied by an undetermined overall phase offset (independent of baseline). This phase offset effectively mixes *Q* and *U*, and must be measured to

obtain a clean separation of CMB power into *E*- and *B*-modes. Calibration of the phase offset requires a source whose polarization angle is known, and we create one by observing RCW38 through polarizing wire grids attached to DASI's 13 receivers. From the wire-grid observations, we can derive the phase offset in each frequency band with an uncertainty of $\lesssim 0.4^\circ$.

As an independent check of this phase offset calibration, the Moon was observed at several epochs during 2001–02. Although the expected amplitude of the polarized signal from the Moon is not well known at these frequencies, the polarization pattern is expected to be radial to high accuracy, and this can be used to determine the cross-polar phase offset independently of the wire grid observations. We show in ref. 51 that these two determinations of the phase offset are in excellent agreement.

On-axis leakage

For ideal polarizers, the cross-polar visibilities are strictly proportional to linear combinations of the Stokes parameters *Q* and *U*. For realistic polarizers, however, imperfect rejection of the unwanted polarization state leads to additional terms in the cross-polar visibilities proportional to the total intensity *I*. These leakage terms are the sum of the complex leakage of the two antennas which form each baseline. During the 2000–01 austral summer, DASI's 13 receivers were retrofitted with multi-element broadband circular polarizers designed to reject the unwanted polarization state to high precision across DASI's 26–36 GHz frequency band. Before installation on the telescope, the polarizers were tuned to minimize these leakages.

At several epochs during 2001–02, the antenna-based leakages were determined with a fractional error of 0.3% from deep observations of the calibrator source RCW38. We show in ref. 51 that antenna-based leakages are $\lesssim 1\%$ (of *I*) at all frequency bands except the highest, for which they approach 2%; this performance is close to the theoretical minimum for our polarizer design. Comparison of the measurements from three epochs separated by many months indicates that the leakages are stable with time.

Given the low level of DASI's leakages, the mixing of power from temperature into polarization in the uncorrected visibilities is expected to be a minor effect at most (see the 'Systematic uncertainties' section). Nonetheless, in the analysis presented in this paper, the cross-polar data have in all cases been corrected to remove this effect using the leakages determined from RCW38.

Off-axis leakage

In addition to on-axis leakage from the polarizers, the feeds will contribute an instrumental polarization that varies across the primary beam. Offset measurements of RCW38 and the Moon indicate that the off-axis instrumental polarization pattern is radial, rising from zero at the beam centre to a maximum of about 0.7% near 3° , and then tapering to zero (see also ref. 51).

With the on-axis polarizer leakage subtracted to $\lesssim 0.3\%$ (see above), this residual leakage, while still quite small compared to the expected level of polarized CMB signal (again, see the 'Systematic uncertainties' section), is the dominant instrumental contribution. Although the visibilities cannot be individually corrected to remove this effect (as for the on-axis leakage), it may be incorporated in the analysis of the CMB data. Using fits to the offset data (see ref. 51 for details), we account for this effect by modelling the contribution of the off-axis leakage to the signal covariance matrix as described in the 'Theory covariance matrix' subsection.

CMB observations and data reduction

Observations

For the observations presented here, two fields separated by 1 h of right ascension, at RA = 23 h 30 min and RA = 00 h 30 min with declination -55° , were tracked continuously. The telescope alternated between the fields every hour, tracking them over precisely the

same azimuth range so that any terrestrial signal can be removed by differencing. Each 24-h period included 20 h of CMB observations and 2.3 h of bracketing calibrator observations, with the remaining time spent on skydips and miscellaneous calibration tasks.

The fields were selected from the 32 fields previously observed by DASI for the absence of any detectable point sources (see ref. 49). The locations of the 32 fields were originally selected to lie at high elevation angle and to coincide with low emission in the IRAS 100 μm and 408 MHz maps of the sky⁵².

The data presented in this paper were acquired from 10 April to 27 October 2001, and again from 14 February to 11 July 2002. In all, we obtained 162 days of data in 2001, and 109 days in 2002, for a total of 271 days before the cuts described in the next section.

Data cuts

Observations are excluded from the analysis, or cut, if they are considered suspect owing to hardware problems, inadequate calibration, contamination from Moon or Sun signal, poor weather or similar effects. In the 'Data consistency and χ^2 tests' section, we describe consistency statistics that are much more sensitive to these effects than are the actual results of our likelihood analysis, allowing us to be certain that the final results are unaffected by contamination. Here we briefly summarize the principal categories of data cuts; we describe each cut in detail in ref. 51.

In the first category of cuts, we reject visibilities for which monitoring data from the telescope indicate obvious hardware malfunction, or simply non-ideal conditions. These include cryogenics failure, loss of tuning for a receiver, large offsets between real/imaginary multipliers in the correlators, and mechanical glitches in the polarizer stepper motors. All data are rejected for a correlator when it shows evidence for large offsets, or excessive noise. An additional cut, and the only one based on the individual data values, is a $>3\sigma$ outlier cut to reject rare ($\ll 0.1\%$ of the data) hardware glitches. Collectively, these cuts reject about 26% of the data.

In the next category, data are cut on the phase and amplitude stability of the calibrator observations. Naturally, we reject data for which bracketing calibrator observations have been lost due to previous cuts. These cuts reject about 5% of the data.

Cuts are also based on the elevation of the Sun and Moon. Co-polar data are cut whenever the Sun was above the horizon, and cross-polar data whenever the solar elevation exceeded 5° . These cuts reject 8% of the data.

An additional cut, which is demonstrably sensitive to poor weather, is based on the significance of data correlations as discussed in the 'Noise model' subsection. An entire day is cut if the maximum off-diagonal correlation coefficient in the data correlation matrix exceeds 8σ significance, referred to gaussian uncorrelated noise. A total of 22 days are cut by this test in addition to those rejected by the solar and lunar cuts.

Reduction

Data reduction consists of a series of steps to calibrate and reduce the data set to a manageable size for the likelihood analysis. Phase and amplitude calibrations are applied to the data on the basis of the bracketing observations of our primary celestial calibrator, RCW38. The raw 8.4-s integrations are combined over each 1-h observation for each of 6,240 visibilities (78 complex baselines $\times$ 10 frequency bands $\times$ 4 Stokes states). In all cases, on-axis leakage corrections are applied to the data, and sequential 1-h observations of the two fields in the same 15° azimuth range are differenced to remove any common ground signal, using a normalization $(\text{field1} - \text{field2})/\sqrt{2}$ that preserves the variance of the sky signal. Except in the case where the data set is split for use in the χ^2 consistency tests in the ' χ^2 tests' subsection, observations from different faceplate rotation angles, epochs and azimuth ranges are all combined, as well as the two co-polar Stokes states, LL and RR. The resulting data set has $N \leq 4,680$ elements ($6,240 \times 3/4 = 4,680$,

where the 3/4 results from the combination of LL and RR). We call this the uncompressed data set, and it contains all of the information in our observations of the differenced fields for Stokes parameters I , Q and U .

Data consistency and χ^2 tests

We begin our analysis by arranging the data into a vector, considered to be the sum of actual sky signal and instrumental noise: $\Delta = \mathbf{s} + \mathbf{n}$. The noise vector $\mathbf{n}$ is hypothesized to be gaussian and random, with zero mean, so that the noise model is completely specified by a known covariance matrix $C_N \equiv \langle \mathbf{n}\mathbf{n}^T \rangle$. Any significant excess variance observed in the data vector Δ will be interpreted as signal. In the likelihood analysis of the next section, we characterize the total covariance of the data set $C = C_T(\kappa) + C_N$ in terms of parameters κ that specify the covariance C_T of this sky signal. This is the conventional approach to CMB data analysis, and it is clear that for it to succeed, the assumptions about the noise model and the accuracy of the noise covariance matrix must be thoroughly tested. This is especially true for our data set, for which long integrations have been used to reach unprecedented levels of sensitivity in an attempt to measure the very small signal covariances expected from the polarization of the CMB.

Noise model

The DASI instrument and observing strategy are designed to remove systematic errors through multiple levels of differencing. Slow and fast phase switching as well as field differencing are used to minimize potentially variable systematic offsets that could otherwise contribute a non-thermal component to the noise. The observing strategy also includes Walsh sequencing of the Stokes states, observations over multiple azimuth ranges and faceplate rotation angles, and repeated observations of the same visibilities on the sky throughout the observing run to allow checks for systematic offsets and verification that the sky signal is repeatable. We hypothesize that after the cuts described in the previous section, the noise in the remaining data is gaussian and white, with no noise correlations between different baselines, frequency bands, real/imaginary pairs, or Stokes states. We have carefully tested the noise properties of the data to validate the use of this model.

Noise variance in the combined data vector is estimated by calculating the variance in the 8.4-s integrations over the period of 1 h, before field differencing. To test that this noise estimate is accurate, we compare three different short-timescale noise estimates: calculated from the 8.4-s integrations over the 1-h observations before and after field differencing and from sequential pairs of 8.4-s integrations. We find that all three agree within 0.06% for co-polar data and 0.03% for cross-polar data, averaged over all visibilities after data cuts.

We also compare the noise estimates based on the short-timescale noise to the variance of the 1-h binned visibilities over the entire data set (up to 2,700 1-hour observations, over a period spanning 457 days). The ratio of long-timescale to short-timescale noise variance, averaged over all combined visibilities after data cuts, is 1.003 for the co-polar data and 1.005 for the cross-polar data, remarkably close to unity. Together with the results of the χ^2 consistency tests described in the ' χ^2 tests' subsection, these results demonstrate that the noise is white and integrates down from timescales of a few seconds to thousands of hours. We find that scaling the diagonal noise by 1% makes a negligible difference in the reported likelihood results (see the 'Systematic uncertainties' section).

To test for potential off-diagonal correlations in the noise, we calculate a $6,240 \times 6,240$ correlation coefficient matrix from the 8.4-s integrations for each day of observations. To increase our sensitivity to correlated noise, we use only data obtained simultaneously for a given pair of data vector elements. Owing to the variable number of 8.4-s integrations M used to calculate each off-

diagonal element, we assess the significance of the correlation coefficient in units of $\sigma = 1/\sqrt{M-1}$. Our weather cut statistic is the daily maximum off-diagonal correlation coefficient significance (see 'Data cuts' subsection).

We use the mean data correlation coefficient matrix over all days, after weather cuts, to test for significant correlations over the entire data set. We find that 1,864 (0.016%) of the off-diagonal elements exceed a significance of 5.5σ , when about one such event is expected for uncorrelated gaussian noise. The outliers are dominated by correlations between real/imaginary pairs of the same baseline, frequency band, and Stokes state, and between different frequency bands of the same baseline and Stokes state. For the real/imaginary pairs, the maximum correlation coefficient amplitude is 0.14, with an estimated mean amplitude of 0.02; for interband correlations the maximum amplitude and estimated mean are 0.04 and 0.003, respectively. We have tested the inclusion of these correlations in the likelihood analysis and find that they have a negligible impact on the results, see the 'Systematic uncertainties' section.

χ^2 tests

As a simple and sensitive test of data consistency, we construct a χ^2 statistic from various splits and subsets of the visibility data. Splitting the data into two sets of observations that should measure the same sky signal, we form the statistic for both the sum and difference data vectors, $\chi^2 = \Delta^T C_N^{-1} \Delta$, where $\Delta = (\Delta_1 \pm \Delta_2)/2$ is the sum or difference data vector, and $C_N = (C_{N1} + C_{N2})/4$ is the corresponding noise covariance matrix. We use the difference data vector, with the common sky signal component subtracted, to test for systematic offsets and mis-estimates of the noise. The sum data vector is used to test for the presence of a sky signal in a straightforward way that is independent of the likelihood analyses that will be used to parameterize and constrain that signal.

We split the data for the difference and sum data vectors in five different ways:

- (1) Year—2001 data versus 2002 data;
- (2) Epoch—the first half of observations of a given visibility versus the second half;
- (3) Azimuth range—east five versus west five observation azimuth ranges;
- (4) Faceplate position—observations at a faceplate rotation angle of 0° versus a rotation angle of 60° ; and
- (5) Stokes state—co-polar observations in which both polarizers are observing left circularly polarized light (LL Stokes state) versus those in which both are observing right circularly polarized light (RR Stokes state).

These splits were done on the combined 2001–02 data set and (except for the first split type) on 2001 and 2002 data sets separately, to test for persistent trends or obvious differences between the years. The faceplate position split is particularly powerful, since the six-fold symmetry of the (u, v) plane coverage allows us to measure a sky signal for a given baseline with a different pair of receivers, different backend hardware, and at a different position on the faceplate with respect to the ground shields, and is therefore sensitive to calibration and other offsets that may depend on these factors. The co-polar split tests the amplitude and phase calibration between polarizer states, and tests for the presence of circularly polarized signal.

For each of these splits, different subsets can be examined: co-polar data only, cross-polar data only (for all except the Stokes state split), various l -ranges (as determined by baseline length in units of wavelength), and subsets formed from any of these which isolate modes with the highest expected signal to noise (s/n). In constructing the high s/n subsets, we must assume a particular theoretical signal template in order to define the s/n eigenmode basis⁵³ appropriate for that subset. For this we use the concordance model defined in the 'Likelihood parameters' subsection, although we find the results are not strongly dependent on the choice of

model. Note that the definitions of which modes are included in the high s/n subsets are made in terms of average theoretical signal, without any reference to the actual data. In Table 1, we present the difference and sum χ^2 values for a representative selection of splits and subsets. In each case we give the degrees of freedom, χ^2 value, and probability to exceed (PTE) this value in the χ^2 cumulative distribution function. For the 296 different split/subset combinations that were examined, the χ^2 values for the difference data appear consistent with noise; among these 296 difference data χ^2 values, there are two with a PTE < 0.01 (the lowest is 0.003), one with a PTE > 0.99 , and the rest appear uniformly distributed between this range. There are no apparent trends or outliers among the various subsets or splits.

The high s/n mode subsets are more sensitive to certain classes of systematic effects in the difference data vector and more sensitive to the expected sky signal in the sum data vector, that otherwise may be masked by noise. Also, the number of modes with expected $s/n > 1$ gives an indication of the power of the experiment to constrain the sky signal. The co-polar data, which are sensitive to the temperature signal, have many more high s/n modes than the cross-polar data,

which measure polarized radiation. Within the context of the concordance model used to generate the s/n eigenmode basis, we have sensitivity with an expected $s/n > 1$ to ~ 340 temperature (co-polar) modes versus ~ 34 polarization (cross-polar) modes.

Detection of signal

Given that the data show remarkable consistency in χ^2 tests of the difference data vectors, the χ^2 values of the sum data vectors can be used to test for the presence of sky signal, independently of the likelihood analysis methods described below. In the co-polar data, all splits and subsets show highly significant χ^2 values (PTE $< 1 \times 10^{-16}$, the precision to which we calculate the cumulative distribution function).

For the cross-polar data, the sum data vector χ^2 values for the high s/n subsets show high significance, with the PTE $< 1 \times 10^{-6}$ for all $s/n > 1$ subsets in Table 1. This simple and powerful test indicates that we have detected, with high significance, the presence of a polarized signal in the data, and that this signal is repeatable in all of the data splits. The polarization map shown in Fig. 1 gives a visual representation of this repeatable polarization signal. Shown are the epoch split sum and difference polarization maps, con-

Table 1 χ^2 Consistency tests for a selection of data splits and subsets

Temperature data						
Split type	Subset	Difference			Sum	
		d.f.	χ^2	PTE	χ^2	PTE
Year	Full	1,448	1,474.2	0.31	23,188.7	$< 1 \times 10^{-16}$
	$s/n > 1$	320	337.1	0.24	21,932.2	$< 1 \times 10^{-16}$
	l range 0–245	184	202.6	0.17	10,566.3	$< 1 \times 10^{-16}$
	l range 0–245 high s/n	36	38.2	0.37	10,355.1	$< 1 \times 10^{-16}$
	l range 245–420	398	389.7	0.61	7,676.0	$< 1 \times 10^{-16}$
	l range 245–420 high s/n	79	88.9	0.21	7,294.4	$< 1 \times 10^{-16}$
	l range 420–596	422	410.5	0.65	3,122.5	$< 1 \times 10^{-16}$
	l range 420–596 high s/n	84	73.5	0.79	2,727.8	$< 1 \times 10^{-16}$
	l range 596–772	336	367.8	0.11	1,379.5	$< 1 \times 10^{-16}$
	l range 596–772 high s/n	67	82.3	0.10	991.8	$< 1 \times 10^{-16}$
	l range 772–1100	108	103.7	0.60	444.4	$< 1 \times 10^{-16}$
	l range 772–1100 high s/n	21	22.2	0.39	307.7	$< 1 \times 10^{-16}$
	Epoch	1,520	1,546.3	0.31	32,767.2	$< 1 \times 10^{-16}$
	$s/n > 1$	348	366.5	0.24	31,430.0	$< 1 \times 10^{-16}$
Azimuth range	Full	1,520	1,542.6	0.34	32,763.8	$< 1 \times 10^{-16}$
	$s/n > 1$	348	355.2	0.38	31,426.9	$< 1 \times 10^{-16}$
Faceplate position	Full	1,318	1,415.2	0.03	27,446.5	$< 1 \times 10^{-16}$
	$s/n > 1$	331	365.3	0.09	26,270.1	$< 1 \times 10^{-16}$
Stokes state	Full	1,524	1,556.6	0.27	33,050.6	$< 1 \times 10^{-16}$
	$s/n > 1$	350	358.2	0.37	31,722.5	$< 1 \times 10^{-16}$
Polarization data						
Split type	Subset	Difference			Sum	
		d.f.	χ^2	PTE	χ^2	PTE
Year	Full	2,896	2,949.4	0.24	2,925.2	0.35
	$s/n > 1$	30	34.4	0.27	82.4	8.7×10^{-7}
	l range 0–245	368	385.9	0.25	315.0	0.98
	l range 0–245 high s/n	73	61.0	0.84	64.5	0.75
	l range 245–420	796	862.2	0.05	829.4	0.20
	l range 245–420 high s/n	159	176.0	0.17	223.8	5.4×10^{-4}
	l range 420–596	844	861.0	0.33	837.3	0.56
	l range 420–596 high s/n	168	181.3	0.23	189.7	0.12
	l range 596–772	672	648.1	0.74	704.4	0.19
	l range 596–772 high s/n	134	139.5	0.35	160.0	0.06
	l range 772–1100	216	192.3	0.88	239.1	0.13
	l range 772–1100 high s/n	43	32.3	0.88	47.6	0.29
	Epoch	3,040	2,907.1	0.96	3,112.2	0.18
	$s/n > 1$	34	29.2	0.70	98.6	3.3×10^{-8}
Azimuth range	Full	3,040	3,071.1	0.34	3,112.9	0.17
	$s/n > 1$	34	38.7	0.27	98.7	3.3×10^{-8}
Faceplate position	Full	2,636	2,710.4	0.15	2,722.2	0.12
	$s/n > 1$	32	43.6	0.08	97.5	1.6×10^{-8}

Results of χ^2 consistency tests for a representative selection of splits and subsets of the combined 2001–02 data set. Visibility data containing the same sky signal is split to form two data vectors; using the instrument noise model, the χ^2 statistic is then calculated on both the difference and sum data vectors. Also tabulated are the number of degrees of freedom (d.f.), and probability to exceed (PTE) the value in the χ^2 cumulative distribution function, to show the significance of the result (PTE values indicated as $< 1 \times 10^{-16}$ are zero to the precision with which we calculate the χ^2 cumulative distribution function). Difference data χ^2 values test for systematic effects in the data, while comparisons with sum data values test for the presence of a repeatable sky signal. Temperature (co-polar) data are visibility data in which the polarizers from both receivers are in the left (LL Stokes state) or right (RR Stokes state) circularly polarized state; polarization (cross-polar) data are those in which the polarizers are in opposite states (LR or RL Stokes state). The ' $s/n > 1$ ' subset is the subset of s/n eigenmodes > 1 and the ' l range high s/n ' subsets are the 20% highest s/n modes within the given l range. See ' χ^2 tests' section for further description of the data split types and subsets. We have calculated 296 χ^2 values for various split types and subsets, with no obvious trends that would indicate systematic contamination of the data.

structured using the same 34 modes with $s/n > 1$ of the polarization data in the concordance model s/n eigenmode basis that appear in Table 1. The sum map shows a repeatable polarized signal, while the difference map is consistent with instrument noise.

It is possible to calculate a similar χ^2 statistic for the data vector formed from the complete, unsplit data set. Combining all the data without the requirement of forming two equally weighted subsets should yield minimal noise, albeit without an exactly corresponding null (that is, difference) test. Recalculating the s/n eigenmodes for this complete cross-polar data vector gives 36 modes with expected $s/n > 1$, for which $\chi^2 = 98.0$ with a $\text{PTE} = 1.2 \times 10^{-7}$. This significance is similar to those from the sum data vectors under the various splits, which actually divide the data fairly equally and so are nearly optimal. It should be noted that our focus so far has been to test the instrumental noise model, and we have not dealt with the small cross-polar signal expected as a result of the off-axis leakage. As noted in the ‘Off-axis leakage’ subsection, it is not possible to correct the data elements directly for this effect, but we can account for it in calculating these χ^2 results by including the expected

covariance of this leakage signal (see ‘Theory covariance matrix’ subsection) in the noise matrix C_N . Again recalculating the s/n eigenmodes, we find 34 cross-polar modes with $s/n > 1$ which give a $\chi^2 = 97.0$ and a $\text{PTE} = 5.7 \times 10^{-8}$, a significance similar to before. The off-axis leakage is also included in the likelihood analyses, where it is again found to have an insignificant impact on the results.

The likelihood analysis described in the following sections makes use of all of the information in our data set. Such an analysis, in principle, may yield statistically significant evidence of a signal even in cases of data sets for which it is not possible to isolate any individual modes which have average $s/n > 1$. However, the existence of such modes in our data set, which has resulted from our strategy of integrating deeply on a limited patch of sky, allows us to determine the presence of the signal with the very simple analysis described above. It also reduces sensitivity to the noise model estimation in the likelihood results that we report next. Finally, it gives our data set greater power to exclude the possibility of no signal than it might have had if we had observed more modes but with less s/n in each.

Likelihood analysis formalism

The preceding section gives strong evidence for the presence of a signal in our polarization data. We now examine that signal using the standard tool of likelihood analysis. In such an analysis, the covariance of the signal, $C_T(\kappa)$, is modelled in terms of parameters κ appropriate for describing the temperature and polarization anisotropy of the CMB. The covariance of the data vector is modelled $C(\kappa) \equiv C_T(\kappa) + C_N$, where C_N is the noise covariance matrix. Given our data vector Δ , the likelihood of the model specified by the parameter vector κ is the probability of our data vector given that model:

$$L(\kappa) = P(\Delta | \kappa) \propto \det(C(\kappa))^{-1/2} \exp\left(-\frac{1}{2} \Delta^T C(\kappa)^{-1} \Delta\right) \quad (1)$$

Although the full likelihood function itself is the most basic result of the likelihood analysis, it is useful to identify and report the values of the parameters that maximize the likelihood (so-called maximum likelihood (ML) estimators). Uncertainties in the parameter values can be estimated by characterizing the shape of the likelihood surface, as discussed further in the ‘Reporting of likelihood results’ subsection.

The CMB power spectra

The temperature and polarization anisotropy of the CMB can be characterized statistically by six angular power spectra: three that give the amplitudes of temperature, E -mode and B -mode polarization anisotropy as a function of angular scale, and three that describe correlations between them. These spectra are written C_l^X , with $X = \{T, E, B, TE, TB, EB\}$. In our likelihood analyses, we choose various parameterizations of these spectra to constrain.

For a given cosmological model, these spectra can be readily calculated using efficient, publicly available Boltzmann codes⁵⁴. Details of how to define these spectra in terms of all-sky multipole expansions of the temperature and linear polarization of the CMB radiation field are available in the literature (see refs 15 and 16). For DAS’s 3.4° field of view, a flat sky approximation is appropriate⁵⁵, so that the spectra may be defined somewhat more simply²⁶. In this approximation the temperature angular power spectrum is defined:

$$C_l^T \approx C^T(u) \equiv \left\langle \frac{\tilde{T}^*(\mathbf{u}) \tilde{T}(\mathbf{u})}{T_{\text{CMB}}^2} \right\rangle \quad (2)$$

where $\tilde{T}(\mathbf{u})$ is the Fourier transform of $T(\mathbf{x})$, T_{CMB} is the mean temperature of the CMB, and $l/2\pi = u$ gives the correspondence between multipole l and Fourier radius $u = |\mathbf{u}|$. The other spectra in

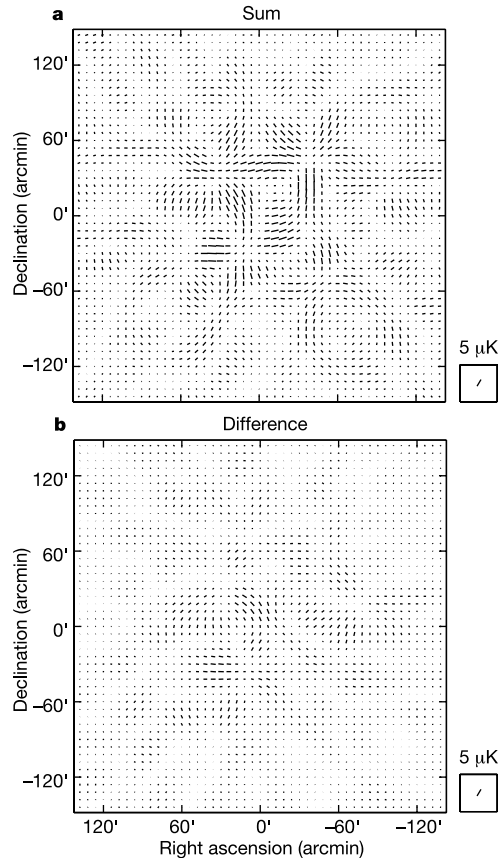


Figure 1 Polarization maps formed from high signal/noise eigenmodes. Shown are maps constructed from polarized data sets that have been split by epoch, and formed into sum (a) or difference (b) data vectors, as reported in section ‘ χ^2 tests’. In order to isolate the most significant signal in our data, we have used only the subset of 34 eigenmodes which, under the concordance model, are expected to have average signal/noise (s/n) > 1 . Because the maps have only 34 independent modes, they exhibit a limited range of morphologies, and unlike conventional interferometer maps, these s/n selected eigenmodes reflect the taper of the primary beam, even when no signal is present. This is visually apparent in the difference map (b), which is statistically consistent with noise. Comparison of the difference map to the sum map (a) illustrates a result also given numerically for this split/subset in Table 1: that these individual modes in the polarized data set show a significant signal.

the flat sky approximation are similarly defined, for example, $C^{TE}(u) \equiv \langle \tilde{T}^*(\mathbf{u}) \tilde{E}(\mathbf{u}) / T_{\text{CMB}}^2 \rangle$. The relationship between $\tilde{E}$, $\tilde{B}$ and the linear polarization Stokes parameters Q and U is:

$$\begin{aligned}\tilde{Q}(\mathbf{u}) &= \cos(2\chi) \tilde{E}(\mathbf{u}) - \sin(2\chi) \tilde{B}(\mathbf{u}) \\ \tilde{U}(\mathbf{u}) &= \sin(2\chi) \tilde{E}(\mathbf{u}) + \cos(2\chi) \tilde{B}(\mathbf{u})\end{aligned}\quad (3)$$

where $\chi = \arg(\mathbf{u})$ and the polarization orientation angle defining Q and U are both measured on the sky from north through east.

Theory covariance matrix

The theory covariance matrix is the expected covariance of the signal component of the data vector, $\mathbf{C}_T \equiv \langle \mathbf{s} \mathbf{s}^T \rangle$. The signals measured by the visibilities in our data vector for a given baseline $\mathbf{u}_i$ (after calibration and leakage correction) are:

$$\begin{aligned}V^{\text{RR}}(\mathbf{u}_i) &= \alpha_i \int d\mathbf{x} A(\mathbf{x}, \nu_i) [T(\mathbf{x}) + V(\mathbf{x})] e^{-2\pi i \mathbf{u}_i \cdot \mathbf{x}} \\ V^{\text{LL}}(\mathbf{u}_i) &= \alpha_i \int d\mathbf{x} A(\mathbf{x}, \nu_i) [T(\mathbf{x}) - V(\mathbf{x})] e^{-2\pi i \mathbf{u}_i \cdot \mathbf{x}} \\ V^{\text{RL}}(\mathbf{u}_i) &= \alpha_i \int d\mathbf{x} A(\mathbf{x}, \nu_i) [Q(\mathbf{x}) + iU(\mathbf{x})] e^{-2\pi i \mathbf{u}_i \cdot \mathbf{x}} \\ V^{\text{LR}}(\mathbf{u}_i) &= \alpha_i \int d\mathbf{x} A(\mathbf{x}, \nu_i) [Q(\mathbf{x}) - iU(\mathbf{x})] e^{-2\pi i \mathbf{u}_i \cdot \mathbf{x}}\end{aligned}\quad (4)$$

where $A(\mathbf{x}, \nu_i)$ specifies the beam power pattern at frequency ν_i , $T(\mathbf{x})$, $Q(\mathbf{x})$, $U(\mathbf{x})$, and $V(\mathbf{x})$ are the four Stokes parameters in units of CMB temperature (μK), and $\alpha_i = \partial B_{\text{Planck}}(\nu_i, T_{\text{CMB}}) / \partial T$ is the appropriate factor for converting from these units to flux density (Jy). The co-polar visibilities V^{RR} and V^{LL} are sensitive to the Fourier transform of the temperature signal $T(\mathbf{x})$ and circular polarization component $V(\mathbf{x})$ (expected to be zero). The cross-polar visibilities V^{RL} and V^{LR} are sensitive to the Fourier transform of the linear polarization components Q and U . Using equation (3), it can be seen that pairwise combinations of the visibilities are direct measures of nearly pure T , E and B Fourier modes on the sky, so that the data set easily lends itself to placing independent constraints on these power spectra.

We construct the theory covariance matrix as the sum of components for each parameter in the analysis:

$$\mathbf{C}_T(\kappa) = \sum_p \kappa_p \mathbf{B}_T^p \quad (5)$$

From equations (2)–(4), it is possible to derive a general expression for the matrix elements of a theory matrix component:

$$\begin{aligned}B_{Tij}^p &= \frac{1}{2} \alpha_i \alpha_j T_{\text{CMB}}^2 \int d\mathbf{u} C^X(u) \tilde{A}(\mathbf{u}_i - \mathbf{u}, \nu_i) \times [\zeta_1 \tilde{A}(\mathbf{u}_j - \mathbf{u}, \nu_j) \\ &\quad + \zeta_2 \tilde{A}(\mathbf{u}_j + \mathbf{u}, \nu_j)]\end{aligned}\quad (6)$$

The coefficients ζ_1 and ζ_2 can take values $\{0, \pm 1, \pm 2\} \times \{\cos\{2\chi, 4\chi\}, \sin\{2\chi, 4\chi\}\}$ depending on the Stokes states (RR, LL, RL, LR) of each of the two baselines i and j and on which of the six spectra (T , E , B , TE , TB , EB) is specified by X . The integration may be limited to annular regions which correspond to l -ranges over which the power spectrum C^X is hypothesized to be relatively flat, or else some shape of the spectrum may be postulated.

Potentially contaminated modes in the data vector may be effectively projected out using a constraint matrix formalism⁵³. This formalism can be used to remove the effect of point sources of known position without knowledge of their flux densities, as we described in ref. 6. This procedure can be generalized to include the case of polarized point sources. Although we have tested for the presence of point sources in the polarization power spectra using this method, in the final analysis we use constraint matrices to project point sources out of the temperature data only, and not the

polarization data (see ‘Point sources’ subsection).

The off-axis leakage, discussed in the ‘Off-axis leakage’ subsection and in detail in ref. 51, has the effect of mixing some power from the temperature signal T into the cross-polar visibilities. Our model of the off-axis leakage allows us to write an expression for it analogous to equation (4), and to construct a corresponding theory covariance matrix component to account for it. In practice, this is a small effect, as discussed in the ‘Systematic uncertainties’ section.

Likelihood parameters

In the ‘Likelihood results’ section we present the results from nine separate likelihood analyses involving the polarization data, the temperature data, or both. Our choice of parameters with which to characterize the six CMB power spectra is a compromise between maximizing sensitivity to the signal and constraining the shape of the power spectra. In the different analyses we either characterize various power spectra with a single amplitude parameter covering all angular scales, or split the l -range into five bands over which spectra are approximated as piecewise-flat, in units of $l(l+1)C_l/(2\pi)$. Five bands were chosen as a compromise between too many for the data to bear and too few to capture the shape of the underlying power spectra. The l -ranges of these five bands are based on those of the nine-band analysis we used in ref. 6; we have simply combined the first four pairs of these bands, and kept the ninth as before. In some analyses we also constrain the frequency spectral indices of the temperature and polarization power spectra as a test for foreground contamination.

The l -range to which DASI has non-zero sensitivity is $28 < l < 1,047$. That range includes the first three peaks of the temperature power spectrum, and within it the amplitude of that spectrum, which we express in units $l(l+1)C_l/(2\pi)$, varies by a factor of about 4. Over this same range, the E -mode polarization spectrum is predicted to have four peaks while rising roughly as l^2 (in the same units), varying in amplitude by nearly two orders of magnitude¹⁷. The TE correlation is predicted to exhibit a complex spectrum that in fact crosses zero five times in this range.

For the single bandpower analyses which maximize our sensitivity to a potential signal, the shape of the model power spectrum assumed will have an effect on the sensitivity of the result. In particular, if the assumed shape is a poor fit to the true spectrum preferred by the data, the results will be both less powerful and difficult to interpret. For temperature spectrum measurements, the most common choice in recent years has been the so-called flat bandpower, $l(l+1)C_l \propto \text{constant}$, which matches the gross large-angle ‘scale-invariant’ power-law shape of that spectrum. Because of extreme variations predicted in the E and TE spectra over DASI’s l -range, we do not expect a single flat bandpower parameterization to offer a good description of the entire data set (although in the ‘ E/B analysis’ subsection we describe results of applying such an analysis to limited l -range subsets of data). A more appropriate definition of ‘flat bandpower’ for polarization measurements sensitive to large ranges of $l < 1,000$ might be $C_l \propto \text{constant}$ (or $l(l+1)C_l \propto l^2$). Other shapes have been tried, notably the gaussian autocorrelation function (by the PIQUE group⁵⁶) which reduces to $C_l \propto \text{constant}$ at large scales and perhaps offers a better fit to the gross shape of the predicted E spectrum.

In our single band analyses, we have chosen a shape for our single bandpower parameters based on the predicted spectra for a cosmological model currently favoured by observations. The specific model that we choose—which we will call the concordance model—is a ΛCDM model with flat spatial curvature, 5% baryonic matter, 35% dark matter, 60% dark energy, and a Hubble constant of $65 \text{ km s}^{-1} \text{ Mpc}^{-1}$, ($\Omega_b = 0.05$, $\Omega_{\text{cdm}} = 0.35$, $\Omega_\Lambda = 0.60$, $h = 0.65$) and the exact normalization $C_{10} = 700 \mu\text{K}^2$. This concordance model was defined in ref. 50 as a good fit to the previous DASI temperature power spectrum and other observations. The concordance model spectra for T , E , and TE are shown in Fig. 4. The five flat

bandpower likelihood results shown in Fig. 4, and discussed in the next section, suggest that the concordance shaped spectra do indeed characterize the data better than any power-law approximation. In the ‘*E/B* analysis’ subsection, we explicitly test the likelihood of the concordance model parameterization against that of the two power laws mentioned above, and find that the concordance model shape is strongly preferred by the data.

It should be noted that the likelihood analysis is always model dependent, regardless of whether a flat or shaped model is chosen for parameterization. To evaluate the expectation value of the results for a hypothesized theoretical power spectrum, we must use window functions appropriate for the parameters of the particular analysis. The calculation of such parameter window functions has previously been described, both generally^{57,58}, and with specific reference to polarization spectra⁵⁹. In general, the parameter window function has a non-trivial shape (even for a flat bandpower analysis) which is dependent on the shape of the true spectra as well as the intrinsic sensitivity of the instrument as a function of angular scale. Parameter window functions for the *E/B* and *E5/B5* polarization analysis are shown in Fig. 2, and will also be made available on our website (<http://astro.uchicago.edu/dasi>).

Likelihood evaluation

Prior to likelihood analysis, the data vector and the covariance matrices can be compressed by combining visibility data from nearby points in the (u, v) plane, where the signal is highly correlated. This reduces the computational time required for the analyses without a significant loss of information about the signal. All analyses were run on standard desktop computers.

For each analysis, we use an iterated quadratic estimator technique to find the ML values of our parameters⁵³. To further characterize the shape of the likelihood function, in ref. 6 we used

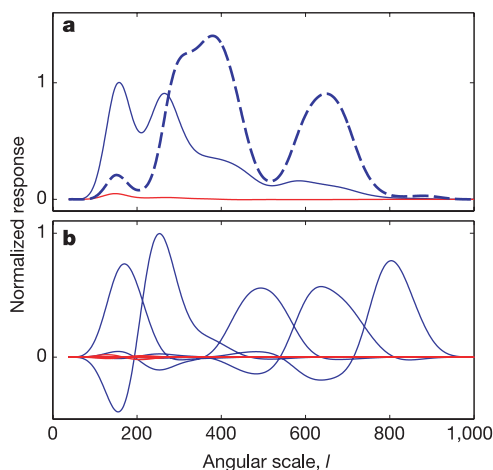


Figure 2 Parameter window functions, which indicate the angular scales over which the parameters in our analyses constrain the power spectra. **a**, The functions for the *E* parameter of our *E/B* analysis, with the solid blue curve indicating response to the *E* power spectrum and the solid red (much lower) curve indicating response of the same *E* parameter to the *B* spectrum. The blue dashed curve shows the result of multiplying the *E* window function by the concordance *E* spectrum, illustrating that for this CMB spectrum, most of the response of our experiment’s *E* parameter comes from the region of the second peak ($300 \lesssim l \lesssim 450$), with a substantial contribution also from the third peak and a smaller contribution from the first. **b**, The *E1* to *E5* parameter window functions for the *E* power spectrum (blue) and *B* power spectrum (red, again much lower) from our *E5/B5* analysis. All of these window functions are calculated with respect to the concordance model discussed in the text. DASI’s response to *E* and *B* is very symmetric, so that *E* and *B* parameter window functions which are calculated with respect to a model for which $E = B$ are nearly identical, with the *E* and *B* spectral response reversed.

an offset log-normal approximation. Here, for improved accuracy in calculating confidence intervals and likelihood ratios, we explicitly map out the likelihood function by evaluating equation (1) over a uniform parameter grid large enough to enclose all regions of substantial likelihood. A single likelihood evaluation typically takes several seconds, so this explicit grid evaluation is impractical for the analyses which include five or more parameters. For each analysis we also implement a Markov chain evaluation of the likelihood function⁶⁰. We find this to be a useful and efficient tool for mapping the likelihoods of these high-dimensional parameter spaces in the region of substantial likelihood. We have compared the Markov technique to the grid evaluation for the lower-dimensional analyses and found the results to be in excellent agreement. In all cases, the peak of the full likelihood evaluated with either technique is confirmed to coincide with the ML values returned by the iterated quadratic estimator.

Simulations and parameter recovery tests

The likelihood analysis software was extensively tested through analysis of simulated data. The analysis software and data simulation software were independently authored, as a check for potential coding errors.

Simulated sky maps were generated from realizations of a variety of smooth CMB power spectra, including both the concordance spectrum and various non-concordance models, both with and without *E* and *B* polarization and *TE* correlations. Independent realizations of the sky were ‘observed’ to construct simulated visibilities with Fourier-plane sampling identical to the real data. The simulations were designed to replicate the actual data as realistically as possible and include artefacts of the instrumental polarization response and calibration, such as the on-axis and off-axis leakages and the cross-polar phase offset, described in the ‘Measuring polarization with DASI’ section, allowing us to test the calibration and treatment of these effects implemented in the analysis software.

Each of the analyses described in the ‘Likelihood results’ section was performed on hundreds of these simulated data sets with independent realizations of sky and instrument noise, both with noise variances that matched the real data, and with noise a factor of ten lower. In all cases, we found that the means of the ML estimators recovered the expectation values $\langle \kappa_p \rangle$ of each parameter without evidence of bias, and that the variance of the ML estimators was found to be consistent with the estimated uncertainty given by F^{-1} evaluated at $\langle \kappa \rangle$, where F is the Fisher matrix.

Reporting of likelihood results

Maximum likelihood (ML) parameter estimates reported in this paper are the global maxima of the multidimensional likelihood function. Confidence intervals for each parameter are determined by integrating (marginalizing) the likelihood function over the other parameters; the reported intervals are the equal-likelihood bounds which enclose 68% of this marginal likelihood distribution. This prescription corresponds to what is generally referred to as the highest posterior density (HPD) interval. When calculating these intervals we consider all parameter values, including non-physical ones, because our aim is simply to summarize the shape of the likelihood function. Values are quoted in the text using the convention ‘ML (HPD-low to HPD-high)’ to make clear that the confidence range is not directly related to the maximum likelihood value.

In the tabulated results, we also report marginalized uncertainties obtained by evaluating the Fisher matrix at the maximum likelihood model, that is, $(F^{-1})_{ii}^{1/2}$ for parameter *i*. Although in most cases, the two confidence intervals are quite similar, we regard the HPD interval as the primary result.

For parameters which are intrinsically positive we consider placing (physical) upper limits by marginalizing the likelihood

distribution as before, but excluding the unphysical negative values. We then test whether the 95% integral point has a likelihood smaller than that at zero; if it does we report an upper limit rather than a confidence interval.

We also report the parameter correlation matrices for our various likelihood analyses to allow the reader to gauge the degree to which each parameter has been determined independently. The covariance matrix is the inverse of the Fisher matrix and the correlation matrix, R , is defined as the covariance matrix normalized to unity on the diagonal, that is, $C = F^{-1}$ and $R_{ij} = C_{ij}/\sqrt{C_{ii}C_{jj}}$.

Goodness-of-fit tests

Using the likelihood function, we wish to determine if our results are consistent with a given model. For example, we would like to examine the significance of any detections by testing for the level of consistency with zero signal models, and we would like to determine if the polarization data are consistent with predictions of the standard cosmological model. We define as a goodness-of-fit statistic the logarithmic ratio of the maximum of the likelihood to its value for some model $\mathcal{H}_0$ described by parameters κ_0 :

$$\Lambda(\mathcal{H}_0) \equiv -\log \left(\frac{L(\kappa_{\text{ML}})}{L(\kappa_0)} \right)$$

The statistic Λ simply indicates how much the likelihood has fallen from its peak value down to its value at κ_0 . Large values indicate inconsistency of the likelihood result with the model $\mathcal{H}_0$. To assess significance, we perform Monte Carlo (MC) simulations of this statistic under the hypothesis that $\mathcal{H}_0$ is true. From this, we can determine the probability, given $\mathcal{H}_0$ true, to obtain a value of Λ that exceeds the observed value, which we hereafter refer to as PTE.

When considering models which the data indicate to be very unlikely, sufficient sampling of the likelihood statistic becomes computationally prohibitive; our typical MC simulations are limited to only 1,000 realizations. In the limit that the parameter errors are normally distributed, our chosen statistic reduces to $\Lambda = \Delta\chi^2/2$. The integral over the χ^2 distribution is described by an incomplete gamma function;

$$\text{PTE} = \frac{1}{\Gamma(N/2)} \int_{\Lambda}^{\infty} e^{-x} x^{\frac{N}{2}-1} dx$$

where $\Gamma(x)$ is the complete gamma function, and N is the number of parameters. Neither the likelihood function nor the distribution of the ML estimators is, in general, normally distributed, and therefore this approximation must be tested. In all cases where we can compute a meaningful PTE with MC simulations, we have done so and found the results to be in excellent agreement with the analytic approximation. Therefore, we are confident that adopting this approximation is justified. All results for PTE in this paper are calculated using this analytic expression unless otherwise stated.

Likelihood results

We have performed nine separate likelihood analyses to constrain various aspects of the signal in our polarization data, in our temperature data, or in both analysed together. The choice of parameters for these analyses and the conventions used for reporting likelihood results have been discussed in ‘Likelihood parameters’ and ‘Reporting of likelihood results’ subsections. Numerical results for the analyses described in this section are given in Tables 2, 3 and 4.

Polarization data analyses and E and B results

E/B analysis. The E/B analysis uses two single-bandpower parameters to characterize the amplitudes of the E and B polarization spectra. As discussed in the ‘Likelihood parameters’ subsection, this analysis requires a choice of shape for the spectra to be parameterized. DASI has instrumental sensitivity to E and B that is symmetrical and nearly independent. Although the B spectrum is not

expected to have the same shape as the E spectrum, we choose the same shape for both spectra in order to make the analysis also symmetrical.

We first compute the likelihood using a $l(l+1)C_l/2\pi = \text{constant}$ bandpower (commonly referred to as ‘flat’ bandpower) including data only from a limited l range in which DASI has high sensitivity; see Fig. 2. Using the range $300 < l < 450$ which includes 24% of the complete data set, we find the ML at flat bandpower values $E = 26.5 \mu\text{K}^2$ and $B = 0.8 \mu\text{K}^2$. The likelihood falls dramatically for the zero polarization ‘nopol’ model $E = 0, B = 0$. Marginalizing over B , we find $\Lambda(E=0) = 16.9$ which, assuming the uncertainties are normally distributed, corresponds to a PTE of 5.9×10^{-9} or a significance of E detection of 5.8σ . As expected, changing the l range affects the maximum likelihood values and the confidence of detection; for example, shifting the centre of the above l range by ± 25 reduces the confidence of detection to 5.6σ and 4.8σ , respectively. Clearly it is desirable to perform the analysis over the entire l range sampled by DASI using a well motivated bandpower shape for the parameterization.

We considered three a priori shapes to check which is most appropriate for our data: the concordance E spectrum shape (as defined in ‘Likelihood parameters’), and two power law alternatives, $l(l+1)C_l \propto \text{constant}$ (flat) and $l(l+1)C_l \propto l^2$. For each of these three cases, the point at $E = 0, B = 0$ corresponds to the same zero-polarization ‘nopol’ model, so that the likelihood ratios $\Lambda(\text{nopol})$ may be compared directly to assess the relative likelihoods of the best-fit models. For the $l(l+1)C_l \propto \text{constant}$ case, the ML flat bandpower values are $E = 6.8 \mu\text{K}^2$, $B = -0.4 \mu\text{K}^2$, with $\Lambda(\text{nopol}) = 4.34$. For the $l(l+1)C_l \propto l^2$ case, the ML values are $E = 5.1 \mu\text{K}^2$, $B = 1.2 \mu\text{K}^2$ (for $l(l+1)C_l/2\pi$ at $l = 300$), with $\Lambda(\text{nopol}) = 8.48$. For the concordance shape, the ML values are $E = 0.80, B = 0.21$ in units of the concordance E spectrum amplitude, with $\Lambda(\text{nopol}) = 13.76$. The likelihood of the best-fit model in the concordance case is a factor of 200 and 12,000 higher than those of the $l(l+1)C_l \propto l^2$ and $l(l+1)C_l \propto \text{constant}$ cases, respectively, and so compared to the concordance shape either of these is a very poor model for the data. The data clearly prefer the concordance shape, which we therefore use for the E/B and other single bandpower analyses presented in our results tables.

Figure 3 illustrates the result of the E/B concordance shape polarization analysis. The maximum likelihood value of E is 0.80 (0.56 to 1.10 at 68% confidence). For B , the result should clearly be regarded as an upper limit; 95% of the $B > 0$ likelihood (marginalized over E) lies below 0.59.

Figure 2a shows the parameter window functions relevant for this analysis. Note that the E parameter has very little sensitivity to B and vice versa—the purity with which DASI can separate these is remarkable. This is also demonstrated by the low correlation (-0.046) between the E and B parameters (see Table 2).

Assuming that the uncertainties in E and B are normally distributed (‘Goodness-of-fit tests’ section), the likelihood ratio $\Lambda(\text{nopol}) = 13.76$ implies a probability that our data are consistent with the zero-polarization hypothesis of $\text{PTE} = 1.05 \times 10^{-6}$. Our data are highly incompatible with the no-polarization hypothesis. Marginalizing over B , we find $\Lambda(E=0) = 12.1$ corresponding to detection of E -mode polarization at a PTE of 8.46×10^{-7} (or a significance of 4.92σ).

The likelihood ratio for the concordance model gives $\Lambda(E=1, B=0) = 1.23$, for which the Monte Carlo and analytic PTE are both 0.28. We conclude that our data are consistent with the concordance model.

However, given the precision to which the temperature power spectrum of the CMB is currently known, even within the ~ 7 -parameter class of cosmological models often considered, the shape and amplitude of the predicted E -mode spectrum are still somewhat uncertain. To quantify this, we have taken the model grid generated for ref. 50 and calculated the expectation value of the

Table 2 Results of likelihood analyses from polarization data

Analysis	Parameter	$l_{\text{low}}-l_{\text{high}}$	ML est.	$(F^{-1})_{ii}^{1/2}$ error	68% interval			Units
					Lower	Upper	UL (95%)	
E/B	E	–	0.80	± 0.28	0.56	1.10	–	Fraction of concordance E
	B	–	0.21	± 0.18	0.05	0.40	0.59	Fraction of concordance E
$E5/B5$	$E1$	28–245	–0.50	± 0.8	–1.20	1.45	2.38	μK^2
	$E2$	246–420	17.1	± 6.3	11.3	31.2	–	μK^2
	$E3$	421–596	–2.7	± 5.2	–10.0	4.3	24.9	μK^2
	$E4$	597–772	17.5	± 16.0	3.8	40.3	47.2	μK^2
	$E5$	773–1,050	11.4	± 49.0	–32.5	92.3	213.2	μK^2
	$B1$	28–245	–0.65	± 0.65	–1.35	0.52	1.63	μK^2
	$B2$	246–420	1.3	± 2.4	–0.7	5.0	10.0	μK^2
	$B3$	421–596	4.8	± 6.5	–0.6	13.5	17.2	μK^2
	$B4$	597–772	13.0	± 14.9	1.6	31.0	49.1	μK^2
	$B5$	773–1,050	–54.0	± 28.9	–77.7	–4.4	147.4	μK^2
E/β_E	E	–	0.84	± 0.28	0.55	1.08	–	Fraction of concordance E
	β_E	–	0.17	± 1.96	–1.63	1.92	–	Temperature spectral index
Scalar/Tensor	S	–	0.87	± 0.29	0.62	1.18	–	Fraction of concordance S
	T	–	–14.3	± 7.5	–20.4	–3.9	25.4	$T/(S = 1)$

The four corresponding parameter correlation matrices

$E1$	$E2$	$E3$	$E4$	$E5$	$B1$	$B2$	$B3$	$B4$	$B5$	E	B
1	–0.137	0.016	–0.002	0.000	–0.255	0.047	–0.004	0.000	0.000	1	–0.046
	1	–0.117	0.014	–0.002	0.024	–0.078	0.004	0.000	0.000		1
		1	–0.122	0.015	–0.003	0.010	–0.027	0.003	–0.001		
			1	–0.119	0.000	–0.001	0.002	–0.016	0.003	E	β_E
				1	0.000	0.000	0.000	0.002	–0.014	1	–0.046
					1	–0.226	0.022	–0.002	0.000		1
						1	–0.097	0.011	–0.002		
							1	–0.111	0.018	S	T
								1	–0.164	1	–0.339
									1		1

ML est., maximum likelihood estimate. $(F^{-1})_{ii}^{1/2}$, Fisher matrix uncertainty for parameter i is evaluated at ML. UL, upper limit.

shaped band E parameter for each model using the window function shown in Fig. 2. We then take the distribution of these predicted E amplitudes, weighted by the likelihood of the corresponding model given our previous temperature results (using a common calibration uncertainty for the DASI temperature and polarization measurements). This yields a 68% credible interval for the predicted value of the E parameter of 0.90 to 1.11. As illustrated in Fig. 3a, our data are compatible with the expectation for E on the basis of existing knowledge of the temperature spectrum. **E5/B5.** Figure 4a and b show the results of a ten-parameter analysis

characterizing the E and B -mode spectra using five flat bandpowers for each. Figure 2b shows the corresponding parameter window functions. Note the extremely small uncertainty in the measurements of the first bands $E1$ and $B1$.

To test whether these results are consistent with the concordance model, we calculate the expectation value for the nominal concordance model in each of the five bands, yielding $E = (0.8, 14, 13, 37, 16)$ and $B = (0, 0, 0, 0, 0) \mu\text{K}^2$. The likelihood ratio comparing this point in the ten-dimensional parameter space to the maximum gives $\Lambda = 5.1$, which for ten degrees of

Table 3 Results of likelihood analyses from temperature data

Analysis	Parameter	$l_{\text{low}}-l_{\text{high}}$	ML est.	$(F^{-1})_{ii}^{1/2}$ error	68% interval		Units
					Lower	Upper	
T/β_T	T	–	1.19	± 0.11	1.09	1.30	Fraction of concordance T
	β_T	–	–0.01	± 0.12	–0.16	0.14	Temperature spectral index
$T5$	$T1$	28–245	6,510	$\pm 1,610$	5,440	9,630	μK^2
	$T2$	246–420	1,780	± 420	1,480	2,490	μK^2
	$T3$	421–596	2,950	± 540	2,500	3,730	μK^2
	$T4$	597–772	1,910	± 450	1,530	2,590	μK^2
	$T5$	773–1,050	3,810	$\pm 1,210$	3,020	6,070	μK^2

The two corresponding parameter correlation matrices

$T1$	$T2$	$T3$	$T4$	$T5$	T	β_T
1	–0.101	0.004	–0.004	–0.001	1	0.023
	1	–0.092	–0.013	–0.011		1
		1	–0.115	–0.010		
			1	–0.147		
				1		

freedom results in a PTE of 0.42, indicating that our data are consistent with the expected polarization parameterized in this way. The $E5/B5$ results are highly inconsistent with the zero-polarization ‘nopol’ hypothesis, for which $\Lambda = 15.2$ with a $\text{PTE} = 0.00073$. This statistic is considerably weaker than the equivalent one obtained for the single band analysis in the ‘E/B analysis’ section, as expected from the higher number of degrees of freedom in this analysis. In this ten-dimensional space, all possible random deviations from the ‘nopol’ expectation values $E = (0,0,0,0,0)$, $B = (0,0,0,0,0)$ are treated equally in constructing the PTE for our Λ statistic. Imagining the ‘nopol’ hypothesis to be true, it would be far less likely to obtain a result in this large parameter space that is both inconsistent with ‘nopol’ at this level and at the same time is consistent with the concordance model, than it would be to obtain a result that is merely inconsistent with ‘nopol’ in some way at this level. It is the latter probability that is measured by the PTE for our $\Lambda(\text{nopol})$, explaining why this approach to goodness-of-fit weakens upon considering increasing numbers of parameters.

E/β_E . We have performed a two-parameter analysis to determine the amplitude of the E -mode polarization signal as above and the frequency spectral index β_E of this signal relative to CMB (Fig. 5). As expected, the results for the E -mode amplitude are very similar to those for the E/B analysis described in the previous section. The spectral index constraint is not strong; the maximum likelihood value is $\beta_E = 0.17$ (-1.63 to 1.92). The result is nevertheless interesting in the context of ruling out possible foregrounds (see the ‘Diffuse foregrounds’ subsection below).

Scalar/tensor. Predictions exist for the shape of the E and B -mode spectra which would result from primordial gravity waves, also known as tensor perturbations, although their amplitude is not well constrained by theory. In a concordance-type model such tensor polarization spectra are expected to peak at $l \approx 100$. Assuming reasonable priors, current measurements of the temperature spectrum (in which tensor and scalar contributions will be mixed) suggest $T/S < 0.2$ (ref. 61), where this amplitude ratio is defined in

terms of the tensor and scalar contributions to the temperature quadrupole C_2^T . We use the distinct polarization angular power spectra for the scalars (our usual concordance E shape, with $B = 0$) and the tensors (E_T and B_T) as two components of a likelihood analysis to constrain the amplitude parameters of these components. In principle, because the scalar B -mode spectrum is zero, this approach avoids the fundamental sample variance limitations arising from using the temperature spectrum alone. However, the $E5/B5$ analysis (see subsection ‘ $E5/B5$ ’) indicates that we have only upper limits to the E - and B -mode polarization at the angular scales most relevant ($l \lesssim 200$) for the tensor spectra. It is therefore not surprising that our limits on T/S derived from the polarization spectra as reported in Table 2 are quite weak.

Temperature data analyses and results for T spectrum

T/β_T . Results are shown in Fig. 5 for a two-parameter analysis to determine the amplitude and frequency spectral index of the temperature signal. The bandpower shape used is that of the concordance T spectrum, and the amplitude parameter is expressed in units relative to that spectrum. The spectral index is relative to the CMB, so that 0 corresponds to a 2.73-K Planck spectrum. The amplitude of T has a maximum likelihood value of 1.19 (1.09 to 1.30), and the spectral index $\beta_T = -0.01$ (-0.16 to 0.14). Although the uncertainty in the temperature amplitude is dominated by sample variance, the spectral index is limited only by the sensitivity and fractional bandwidth of DASI. Owing to the extremely high s/n of the temperature data, the constraints on the spectral index are superior to those from previous DASI observations (ref. 6).

$T5$. Fig. 4c shows the results of an analysis using five flat bands to characterize the temperature spectrum. These results are completely dominated by the sample variance in the differenced field. They are consistent with, although less precise than our previous temperature power spectra described in ref. 6; we include them here primarily to emphasize that DASI makes measurements simultaneously in all four Stokes parameters and is therefore able to

Table 4 Results of likelihood analyses from joint temperature-polarization data set

Analysis	Parameter	$l_{\text{low}}-l_{\text{high}}$	ML est.	$(F^{-1})_{ii}^{1/2}$ error	68% interval		Units
					Lower	Upper	
$T/E/TE$	T	–	1.13	± 0.10	1.05	1.29	Fraction of concordance T
	E	–	0.77	± 0.27	0.57	1.10	Fraction of concordance E
	TE	–	0.91	± 0.38	0.45	1.37	Fraction of concordance TE
$T/E/TE5$	T	–	1.12	± 0.10	1.09	1.31	Fraction of concordance T
	E	–	0.81	± 0.28	0.71	1.36	Fraction of concordance E
	$TE1$	28–245	–24.8	± 32.2	–55.3	24.7	μK^2
	$TE2$	246–420	92.3	± 38.4	44.9	151.1	μK^2
	$TE3$	421–596	–10.5	± 48.2	–60.1	52.0	μK^2
	$TE4$	597–772	–66.7	± 74.3	–164.6	9.5	μK^2
	$TE5$	773–1050	20.0	± 167.9	–130.3	172.3	μK^2
$T/E/B/TE/TB/EB$	T	–	1.13	± 0.10	1.03	1.27	Fraction of concordance T
	E	–	0.75	± 0.26	0.59	1.19	Fraction of concordance E
	B	–	0.20	± 0.18	0.11	0.52	Fraction of concordance E
	TE	–	1.02	± 0.37	0.53	1.49	Fraction of concordance TE
	TB	–	0.53	± 0.32	0.08	0.82	Fraction of concordance TE
	EB	–	–0.16	± 0.16	–0.38	0.01	Fraction of concordance E

The three parameter correlation matrices

T	E	$TE1$	$TE2$	$TE3$	$TE4$	$TE5$	T	E	B	TE	TB	EB
1	0.026	–0.071	0.202	–0.018	–0.075	0.001	1	0.026	0.004	0.230	0.136	0.033
	1	–0.067	0.339	–0.023	–0.090	0.008		1	–0.027	0.320	–0.040	–0.182
		1	–0.076	0.006	0.011	–0.001			1	–0.027	0.219	–0.190
			1	–0.078	–0.039	0.004				1	–0.150	0.109
				1	–0.056	0.004					1	0.213
					1	–0.066						1
T	E	TE										
1	0.017	0.207										
	1	0.282										
		1										

measure temperature as well as polarization anisotropy. Note that these results and those for T/β_T have not been corrected for residual point sources.

Joint analyses and cross spectra results

$T/E/TE$. Figure 6 shows the results of a three-parameter single bandpower analysis of the amplitudes of the T and E spectra, and the TE cross-correlation spectrum. As before, bandpower shapes based on the concordance model are used. The T and E constraints are, as expected, very similar to those from the E/B , E/β_E and T/β_T analyses described above. The new result here is TE which has a maximum likelihood value of 0.91 (0.45 to 1.37). Note that in contrast to the two-dimensional likelihoods shown in other figures, here the contours show apparent evidence of correlation between the two parameters; the parameter correlation coefficients from Table 4 are 0.28 for E/TE and 0.21 for T/TE .

Marginalizing over T and E , we find that the likelihood of TE peaks very near 1, so that $\Lambda(TE = 1) = 0.02$ with a PTE of 0.857. For the ‘no cross-correlation’ hypothesis, $\Lambda(TE = 0) = 1.85$ with an analytic PTE of 0.054 (the PTE calculated from Monte Carlo simulations is 0.047). This result represents a detection of the expected TE correlation at 95% confidence and is particularly interesting in that it suggests a common origin for the observed temperature and polarization anisotropy.

It has been suggested that an estimator of TE cross-correlation constructed using a $TE = 0$ prior may offer greater immunity to systematic errors⁵⁹. We have confirmed that applying such a technique to our data yields similar results to the above likelihood analysis, with errors slightly increased as expected.

$T/E/TE5$. We have performed a seven-parameter analysis using single shaped bandpowers for T and E , and five flat bandpowers for the TE cross-correlation; the TE results from this are shown in Fig. 4d. In this analysis the B -mode polarization has been explicitly set to zero. Again, the T and E constraints are similar to the values for the other analyses where these parameters appear. The TE

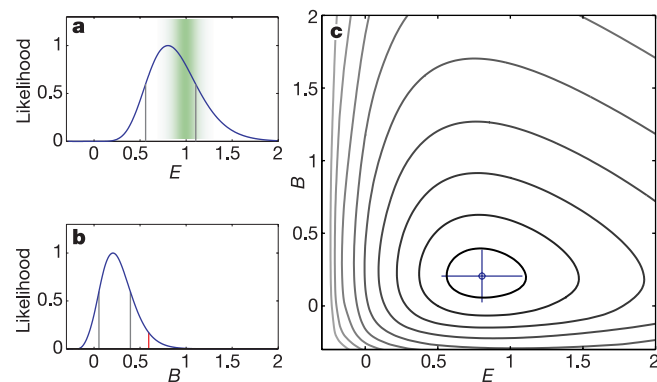


Figure 3 Results from the two-parameter shaped bandpower E/B polarization analysis. An E -mode power spectrum shape as predicted for the concordance model is assumed, and the units of amplitude are relative to that model. The same shape is assumed for the B -mode spectrum. **c.** The point shows the maximum likelihood value with the cross indicating Fisher matrix errors. Likelihood contours are placed at levels $\exp(-n^2/2)$, $n = 1, 2, \dots$, relative to the maximum, that is, for a normal distribution, the extrema of these contours along either dimension would give the marginalized n -sigma interval.

a, b. The corresponding single parameter likelihood distributions marginalized over the other parameter. Note the steep fall in likelihood toward low power values; this likelihood shape (similar to a χ^2 distribution) is typical for positive-definite parameters for which a limited number of high s/n modes are relevant. The grey lines enclose 68% of the total likelihood. The red line indicates the 95% confidence upper limit on B -mode power. The green band shows the distribution of E expectation values for a large grid of cosmological models weighted by the likelihood of those models given our previous temperature result (see ref. 50).

bandpowers are consistent with the predictions of the concordance model.

$T/E/B/TE/TB/EB$. Finally, we describe the results of a six shaped bandpower analysis for the three individual spectra T , E and B , together with the three possible cross-correlation spectra TE , TB and EB . We include the B cross-spectra for completeness, though there is little evidence for any B -mode signal. Because there are no predictions for the shapes of the TB or EB spectra (they are expected to be zero), we preserve the symmetry of the analysis between E and B by simply parameterizing them in terms of the TE and E spectral shapes. The results for T , E , B and TE are similar to those obtained before, with no detection of EB or TB .

Systematic uncertainties

Noise, calibration, offsets and pointing

To assess the effect of systematic uncertainties on the likelihood results, we have repeated each of the nine analyses with alternative assumptions about the various effects that we have identified which reflect the range of uncertainty on each.

Much of the effort of the data analysis presented in this paper has gone into investigating the consistency of the data with the noise model as discussed in the ‘Noise model’ subsection. We find no discrepancies between complementary noise estimates on different timescales, to a level $\ll 1\%$. As discussed in the ‘ χ^2 tests’ subsection, numerous consistency tests on subsets of the co-polar and

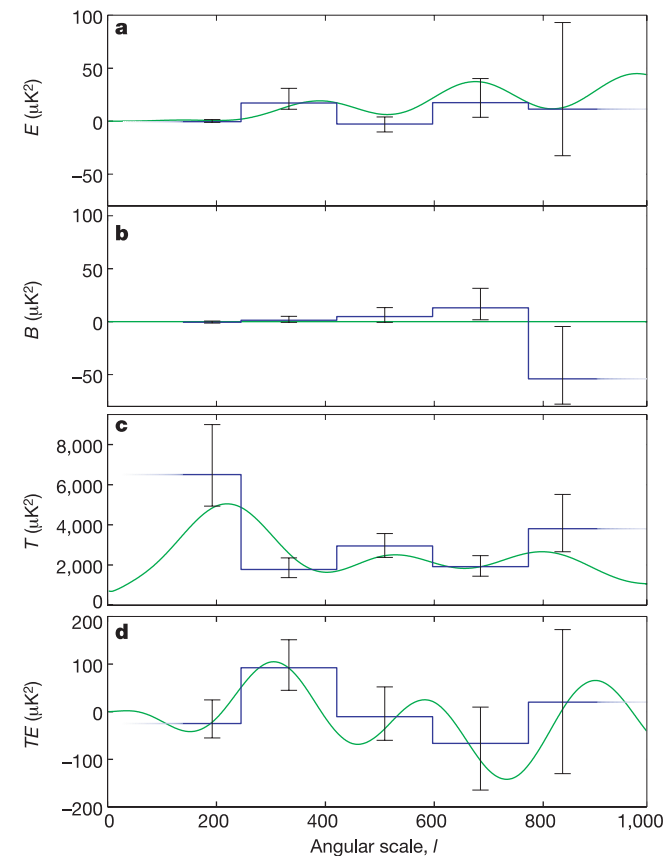


Figure 4 Results from several likelihood analyses. The ten-parameter $E5/B5$ polarization analysis is shown in **a** and **b**. The $T5$ temperature analysis is shown in **c** and the five TE bands from the $T/E/TE5$ joint analysis are shown in **d**. All the results are shown in flat bandpower units of $l(l+1)C_l/(2\pi)$. The blue line shows the piecewise flat bandpower model for the maximum likelihood parameter values, with the error bars indicating the 68% central region of the likelihood of each parameter, marginalizing over the other parameter values (analogous to the grey lines in Fig. 3a and b). In each case the green line is the concordance model.

cross-polar visibility data show no evidence for an error in the noise scaling to a similar level. When we re-evaluate each of the analyses described in the ‘Likelihood results’ section with the noise scaled by 1%, the shift in the maximum likelihood values for all parameters is entirely negligible.

In the ‘Noise model’ subsection, we reported evidence of some detectable noise correlations between real/imaginary visibilities and between visibilities from different bands of the same baseline. When either or both of these noise correlations are added to the covariance matrix at the measured level, the effects are again negligible: the most significant shift is in the highest- l band of the E spectrum from the $E5/B5$ analysis (see the ‘ $E5/B5$ ’ subsection), where the power shifts by about $2\ \mu\text{K}^2$.

Errors in the determination of the absolute cross-polar phase offsets will mix power between E and B ; these phase offsets have been independently determined from wire-grid calibrations and observations of the Moon, and found to agree to within the measurement uncertainties of about 0.4° (ref. 51). Reanalysis of the data with the measured phase offsets shifted by 2° demonstrates that the likelihood results are immune to errors of this magnitude: the most significant effect occurs in the highest- l band of the TE spectrum from the $T, E, TE5$ analysis (see the ‘ $T/E/TE5$ ’ subsection), where the power shifts by about $30\ \mu\text{K}^2$.

The on-axis leakages described in the ‘On-axis leakage’ subsection will mix power from T into E and B , and the data are corrected for this effect in the course of reduction, before input to any analyses. When the likelihood analyses are performed without the leakage correction, the largest effects appear in the shaped TE amplitude analysis (see ‘ $T/E/TE$ ’ subsection), and the lowest- l band of $TE5$ from the $T, E, TE5$ analysis (see the ‘ $T/E/TE5$ ’ subsection); all shifts are tiny compared to the 68% confidence

intervals. As the leakage correction itself has little impact on the results, the uncertainties in the correction, which are at the $< 1\%$ level, will have no noticeable effect.

As described in the ‘Off-axis leakage’ subsection, the off-axis leakage from the feeds is a more significant effect, and is accounted for in the likelihood analysis by modelling its contribution to the covariance matrix. When this correction is not applied, the E, B results (see ‘ E/B analysis’ subsection) shift by about 4% and 2%, respectively, as expected from simulations of this effect. Although this bias is already small, the simulations show that the correction removes it completely to the degree to which we understand the off-axis leakage. Uncertainties in the leakage profiles of the order of the fit residuals (see ref. 51) lead to shifts of less than 1%.

The pointing accuracy of the telescope is measured to be better than 2 arcmin and the root-mean-square tracking errors are < 20 arcsec; as we discussed in refs 49 and 6, this is more than sufficient for the characterization of CMB anisotropy at the much larger angular scales measured by DASI.

Absolute calibration of the telescope was achieved through measurements of external thermal loads, transferred to the calibrator RCW38. The dominant uncertainty in the calibration is due to temperature and coupling of the thermal loads. As reported in ref. 6, we estimate an overall calibration uncertainty of 8% (1σ), expressed as a fractional uncertainty on the C_l bandpowers (4% in $\Delta T/T$). This applies equally to the temperature and polarization data presented here.

Foregrounds

Point sources. The highest-sensitivity point-source catalogue in our observing region is the 5-GHz PMN survey⁶². For our first-season temperature analysis described in refs 49 and 6 we projected out known sources using this catalogue. We have kept this procedure for the temperature data presented here, projecting the same set of sources as before.

Unfortunately the PMN survey is not polarization sensitive. We note that the distribution of point-source polarization fractions is approximately exponential (see below). Total intensity is thus a poor indicator of polarized intensity and it is therefore not sensible to project out the PMN sources in our polarization analysis.

Our polarization fields were selected for the absence of any significant point-source detections in the first-season data. No significant detections are found in the 2001–02 data, either in the

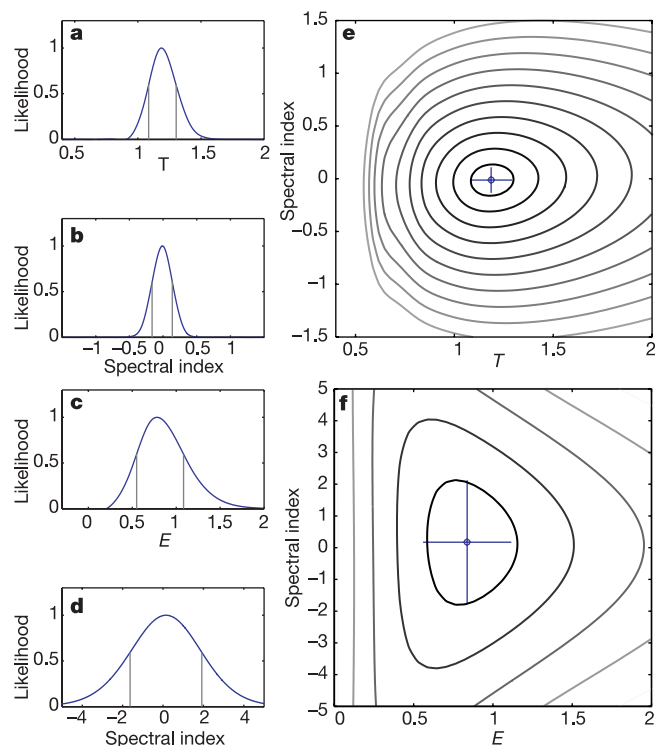


Figure 5 Results of shaped bandpower amplitude/spectral-index analyses. **a, b, e**, The T/β_T temperature analysis assuming the T power spectrum shape as predicted for the concordance model, and in units relative to that model. The layout of the plot is analogous to Fig. 3. Spectral index is relative to the CMB blackbody—in these units, synchrotron emission would be expected to have an index of approximately -3 . **c, d, f**, Results of the similar E/β_E analysis performed on the polarization data.

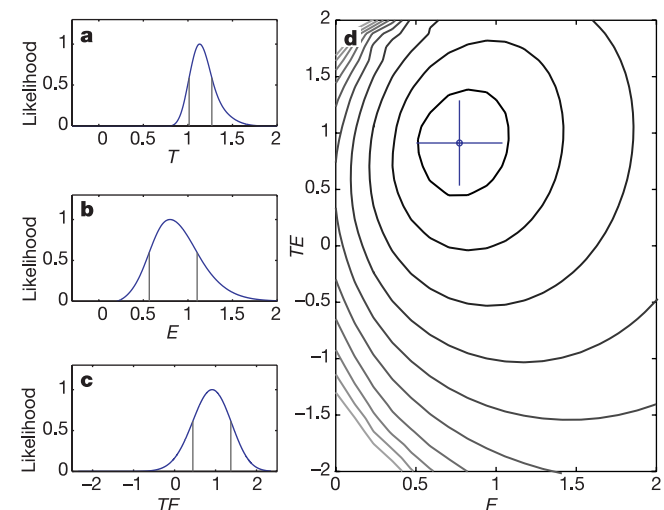


Figure 6 Results from the three-parameter shaped bandpower $T/E/TE$ joint analysis. Spectral shapes as predicted for the concordance model are assumed (**a–c**) and the units are relative to that model. The layout of the plot is analogous to Fig. 3. The two-dimensional distribution in **d** is marginalized over the T dimension.

temperature data, which are dominated by CMB anisotropy, or in the polarization data.

To calculate the expected contribution of undetected point sources to our polarization results we would like to know the distribution of polarized flux densities, but unfortunately no such information exists in our frequency range. However, to make an estimate, we use the distribution of total intensities, and then assume a distribution of polarization fractions. We know the former distribution quite well from our own first-season 32-field data where we detect 31 point sources and determine that $dN/dS_{31} = (32 \pm 7)S^{(-2.15 \pm 0.20)} \text{ Jy}^{-1} \text{ sr}^{-1}$ in the range 0.1 to 10 Jy. This is consistent, when extrapolated to lower flux densities, with a result from the CBI experiment valid in the range 5–50 mJy (ref. 63). The distribution of point source polarization fractions at 5 GHz can be characterized by an exponential with a mean of 3.8% (ref. 64); data of somewhat lower quality at 15 GHz are consistent with the same distribution⁶⁵. Qualitatively, we expect the polarization fraction of synchrotron-dominated sources initially to rise with frequency, and then reach a plateau or fall, with the break point at frequencies $\ll 5$ GHz (see ref. 66 for an example). In the absence of better data we have conservatively assumed that the exponential distribution mentioned above continues to hold at 30 GHz.

We proceed to estimate the effect of point sources by Monte Carlo simulation, generating realizations using the total intensity and polarization fraction distributions mentioned above. For each realization, we generate simulated DASI data by adding a realization of CMB anisotropy and appropriate instrument noise. The simulated data are tested for evidence of point sources and those realizations that show statistics similar to the real data are kept. The effect of off-axis leakage, which we describe and quantify in ref. 51, is included in these calculations.

When the simulated data are passed through the *E/B* analysis described in the ‘*E/B* analysis’ subsection, the mean bias of the *E* parameter is 0.04 with a standard deviation of 0.05; in 95% of cases the shift distance in the *E/B* plane is less than 0.13. We conclude that the presence of point sources consistent with our observed data has a relatively small effect on our polarization results.

Diffuse foregrounds. In ref. 51, we find no evidence for contamination of the temperature data by synchrotron, dust or free-free emission, as confirmed by the limits on the temperature spectral index presented in the ‘*T/β_T*’ subsection. The expected fractional polarization of the CMB is of order 10%, while the corresponding number for free-free emission is less than 1%. Diffuse thermal dust emission may be polarized by several per cent (see for example ref. 67), although we note that polarization of the admixture of dust and free-free emission observed with DASI in NGC 6334 is $\ll 1\%$ (see ref. 51). Likewise, emission from spinning dust is not expected to be polarized at detectable levels⁶⁸. Therefore if free-free and dust emission did not contribute significantly to our temperature anisotropy results they are not expected to contribute to the polarization. Synchrotron emission on the other hand can in principle be up to 70% polarized, and is by far the greatest concern; what was a negligible contribution in the temperature case could be a significant one in polarization.

There are no published polarization maps in the region of our fields. Previous attempts to estimate the angular power spectrum of polarized synchrotron emission have been guided by surveys of the Galactic plane at frequencies of 1–3 GHz (refs 69 and 70). These maps show much more small-scale structure in polarization than in temperature, but this is mostly induced by Faraday rotation⁷¹, an effect which is negligible at 30 GHz. In addition, because synchrotron emission is highly concentrated in the disk of the Galaxy it is not valid to assume that the angular power spectrum at low Galactic latitudes has much to tell us about that at high latitudes⁷².

Our fields lie at Galactic latitude -58.4° and -61.9° . The brightness of the IRAS 100 μm and Haslam 408 MHz (ref. 52) maps within our fields lie at the 6% and 25% points, respectively,

of the integral distributions taken over the whole sky. There are several strong pieces of evidence from the DASI data set itself that the polarization results described in this paper are free of significant synchrotron contamination. The significant *TE* correlation shown in Fig. 6 indicates that the temperature and *E*-mode signal have a common origin. The tight constraints on the temperature anisotropy spectral index require that this common origin has a spectrum consistent with CMB. Galactic synchrotron emission is known to have a temperature spectral index of -2.8 (ref. 73), with evidence for steepening to -3.0 at frequencies above 1–2 GHz (ref. 74). At frequencies where Faraday depolarization is negligible (> 10 GHz), the same index will also apply for polarization. The dramatically tight constraint on the temperature spectral index of $\beta_T = 0.01$ (-0.16 to 0.14) indicates that any component of the temperature signal coming from synchrotron emission is negligibly small in comparison to the CMB. More directly, the constraint on the *E*-mode spectral index $\beta_E = 0.17$ (-1.63 to 1.92) disfavors synchrotron polarization at nearly 2σ . A third, albeit weaker, line of argument is that a complex synchrotron emitting structure is not expected to produce a projected brightness distribution which prefers *E*-mode polarization over *B*-mode⁷⁵. Therefore, the result in Fig. 3 could be taken as further evidence that the signal we are seeing is not due to synchrotron emission.

Discussion

This paper presents several measures of the confidence with which CMB polarization has been detected with DASI. Which measure is preferred depends on the desired level of statistical rigour and independence from a priori models of the polarization. The χ^2 analyses in the ‘ χ^2 tests’ subsection offer the most model-independent results, although the linear combinations of the data used to form the *s/n* eigenmodes are selected by consideration of theory and the noise model. For the high *s/n* eigenmodes of the polarization data, the probability to exceed (PTE) the measured χ^2 for the sum of the various data splits ranges from 1.6×10^{-8} to 8.7×10^{-7} , while the χ^2 for the differences are found to be consistent with noise. The PTE for the χ^2 found for the total (that is, not split) high *s/n* polarization eigenmodes, corrected for the beam offset leakage, is 5.7×10^{-8} .

Likelihood analyses are in principle more model dependent, and the analyses reported make different assumptions for the shape of the polarization power spectrum. Using theory to select the angular scales on which DASI should be most sensitive, we calculate the likelihood for a flat bandpower in *E* and *B* over the multipole range $300 < l < 450$ and find that data are consistent with $B = 0$ over this range, but that $E = 0$ can be rejected with a PTE of 5.9×10^{-9} .

The choice of the model bandpower is more important when the full *l* range of the DASI data is analysed. In this case, the likelihood analyses indicate that a $l(l+1)C_l \propto l^2$ model for the *E*-mode spectrum is 60 times more likely than a $l(l+1)C_l = \text{constant}$ model. Further, a bandpass shape based on the concordance model is found to be 12,000 times more likely than the flat bandpower. The likelihood ratio test leads to a PTE of 8.5×10^{-7} for the concordance shaped bandpower, corresponding to a confidence of detection of 4.9σ . In all three of these tests, *B* is found to be consistent with zero, as expected in the concordance model.

The concordance model is also supported by the results of the five-band piecewise-flat analyses, *E5/B5*. The upper limit for the first *E* band at $28 \leq l \leq 245$ is only $2.38 \mu\text{K}^2$, a factor of 30 lower in power than the previous upper limit. The next band at $246 \leq l \leq 420$ is detected with a maximum likelihood value of $17.1 \mu\text{K}^2$. Such a sharp rise in power with increasing *l* is expected in the *E*-mode spectrum (see Fig. 4) owing to the length scale introduced by the mean free path to photon scattering. The polarization of the larger modes is suppressed because the velocity differences are not as large on the scale of the mean free path. In fact, the *E* spectrum is expected to increase as $l(l+1)C_l \propto l^2$ at small *l*. This dependence is not expected to continue to the higher *l* values to

which DASI is sensitive, owing to diffusion damping, which suppresses power on scales smaller than the mean free path. The maximum likelihood values of the higher l bands of the DASI five-band E analysis are consistent with the damped concordance model, but lie below a simple extrapolation of the l^2 power law. Again, the five-band B -mode spectrum is consistent with the concordance prediction of zero.

The TE analysis provides further confidence in our detection of CMB polarization and for the concordance model. From the $T/E/TE$ likelihood analysis, the $TE = 0$ hypothesis is rejected with a PTE of 0.054. Note that TE could be negative as well as positive and therefore the $TE \leq 0$ hypothesis is rejected with higher confidence.

Lastly, the measured T frequency spectral index, 0.01 (-0.16 to 0.14), is remarkably well constrained to be thermal and is inconsistent with any known foregrounds. The E frequency spectral index, 0.17 (-1.63 to 1.92), is also consistent with the CMB, and although less well constrained than the T index, is inconsistent with diffuse foreground synchrotron emission at nearly 2σ .

In summary, the analyses reported in this paper all indicate a robust detection of E -mode CMB polarization with a confidence level $\geq 4.9\sigma$. The measured properties of the polarization are in good agreement with predictions of the concordance model and, as discussed in the 'Foregrounds' subsection, are inconsistent with expectations from known sources of foreground emission. These results therefore provide strong support for the underlying theoretical framework for the generation of CMB anisotropy. They lend confidence to the values of the cosmological parameters and to the extraordinary picture of the origin and structure of the Universe derived from CMB measurements. The prospect of further refining our understanding of the Universe using precision polarization measurements is the goal of many ongoing and planned CMB experiments. The detection of polarization at the predicted level reported in this paper points to a promising future for the field. $\square$

Received 7 October; accepted 5 November 2002; doi:10.1038/nature01269.

1. Penzias, A. A. & Wilson, R. W. A measurement of excess antenna temperature at 4080 Mc/s. *Astrophys. J.* **142**, 419–421 (1965).
2. Mather, J. C. *et al.* Measurement of the cosmic microwave background spectrum by the COBE FIRAS instrument. *Astrophys. J.* **420**, 439–444 (1994).
3. Fixsen, D. J. *et al.* The cosmic microwave background spectrum from the full COBE FIRAS data set. *Astrophys. J.* **473**, 576–587 (1996).
4. Smoot, G. F. *et al.* Structure in the COBE differential microwave radiometer first-year maps. *Astrophys. J.* **396**, L1–L5 (1992).
5. Miller, A. D. *et al.* A measurement of the angular power spectrum of the cosmic microwave background from $l = 100$ to 400 . *Astrophys. J.* **524**, L1–L4 (1999).
6. Halverson, N. W. *et al.* Degree angular scale interferometer first results: A measurement of the cosmic microwave background angular power spectrum. *Astrophys. J.* **568**, 38–45 (2002).
7. Netterfield, C. B. *et al.* A measurement by BOOMERANG of multiple peaks in the angular power spectrum of the cosmic microwave background. *Astrophys. J.* **571**, 604–614 (2002).
8. Lee, A. T. *et al.* A high spatial resolution analysis of the MAXIMA-1 cosmic microwave background anisotropy data. *Astrophys. J.* **561**, L1–L5 (2001).
9. Pearson, T. J. *et al.* The anisotropy of the microwave background to $l = 3500$: Mosaic observations with the cosmic background imager. *Astrophys. J.* (submitted); preprint astro-ph/0205388 at (<http://xxx.lanl.gov>) (2002).
10. Scott, P. F. *et al.* First results from the Very Small Array - III. The CMB power spectrum. *Mon. Not. R. Astron. Soc.* (submitted); preprint astro-ph/0205380 at (<http://xxx.lanl.gov>) (2002).
11. Hu, W. & Dodelson, S. Cosmic microwave background anisotropies. *Annu. Rev. Astron. Astrophys.* **40**, 171–216 (2002).
12. Kaiser, N. Small-angle anisotropy of the microwave background radiation in the adiabatic theory. *Mon. Not. R. Astron. Soc.* **202**, 1169–1180 (1983).
13. Bond, J. R. & Efstathiou, G. Cosmic background radiation anisotropies in universes dominated by nonbaryonic dark matter. *Astrophys. J.* **285**, L45–L48 (1984).
14. Polnarev, A. G. Polarization and anisotropy induced in the microwave background by cosmological gravitational waves. *Sov. Astron.* **29**, 607–613 (1985).
15. Kamionkowski, M., Kosowsky, A. & Stebbins, A. Statistics of cosmic microwave background polarization. *Phys. Rev. D* **55**, 7368–7388 (1997).
16. Zaldarriaga, M. & Seljak, U. All-sky analysis of polarization in the microwave background. *Phys. Rev. D* **55**, 1830–1840 (1997).
17. Hu, W. & White, M. A CMB polarization primer. *New Astron.* **2**, 323–344 (1997).
18. Kosowsky, A. Introduction to microwave background polarization. *New Astron. Rev.* **43**, 157–168 (1999).
19. Hu, W., Spergel, D. N. & White, M. Distinguishing causal seeds from inflation. *Phys. Rev. D* **55**, 3288–3302 (1997).
20. Kinney, W. H. How to fool cosmic microwave background parameter estimation. *Phys. Rev. D* **63**, 43001 (2001).

21. Bucher, M., Moodley, K. & Turok, N. Constraining isocurvature perturbations with cosmic microwave background polarization. *Phys. Rev. Lett.* **87**, 191301 (2001).
22. Rees, M. J. Polarization and spectrum of the primeval radiation in an anisotropic universe. *Astrophys. J.* **153**, L1–L5 (1968).
23. Zaldarriaga, M. & Harari, D. D. Analytic approach to the polarization of the cosmic microwave background in flat and open universes. *Phys. Rev. D* **52**, 3276–3287 (1995).
24. Coulson, D., Crittenden, R. G. & Turok, N. G. Polarization and anisotropy of the microwave sky. *Phys. Rev. Lett.* **73**, 2390–2393 (1994).
25. Crittenden, R., Davis, R. L. & Steinhardt, P. J. Polarization of the microwave background due to primordial gravitational waves. *Astrophys. J.* **417**, L13–L16 (1993).
26. Seljak, U. Measuring polarization in the cosmic microwave background. *Astrophys. J.* **482**, 6–16 (1997).
27. Kamionkowski, M., Kosowsky, A. & Stebbins, A. A probe of primordial gravity waves and vorticity. *Phys. Rev. Lett.* **78**, 2058–2061 (1997).
28. Seljak, U. & Zaldarriaga, M. Signature of gravity waves in the polarization of the microwave background. *Phys. Rev. Lett.* **78**, 2054–2057 (1997).
29. Lyth, D. H. What would we learn by detecting a gravitational wave signal in the Cosmic Microwave Background anisotropy? *Phys. Rev. Lett.* **78**, 1861–1863 (1997).
30. Zaldarriaga, M. & Seljak, U. Gravitational lensing effect on cosmic microwave background polarization. *Phys. Rev. D* **58**, 23003 (1998).
31. Hu, W. & Okamoto, T. Mass reconstruction with cosmic microwave background polarization. *Astrophys. J.* **574**, 566–574 (2002).
32. Knox, L. & Song, Y. A limit on the detectability of the energy scale of inflation. *Phys. Rev. Lett.* **89**, 011303 (2002).
33. Kesden, M., Cooray, A. & Kamionkowski, M. Separation of gravitational-wave and cosmic-shear contributions to cosmic microwave background polarization. *Phys. Rev. Lett.* **89**, 011304 (2002).
34. Kamionkowski, M. & Kosowsky, A. The cosmic microwave background and particle physics. *Annu. Rev. Nucl. Part. Sci.* **49**, 77–123 (1999).
35. Staggs, S. T., Gunderson, J. O. & Church, S. E. *ASP Conf. Ser.* **181**, *Microwave Foregrounds* (eds de Oliveira-Costa, A. & Tegmark, M.) 299–309 (Astronomical Society of the Pacific, San Francisco, 1999).
36. Caderni, N., Fabbri, R., Melchiorri, B., Melchiorri, F. & Natale, V. Polarization of the microwave background radiation. II. An infrared survey of the sky. *Phys. Rev. D* **17**, 1908–1918 (1978).
37. Nanos, G. P. Polarization of the blackbody radiation at 3.2 centimeters. *Astrophys. J.* **232**, 341–347 (1979).
38. Lubin, P. M. & Smoot, G. F. Search for linear polarization of the cosmic background radiation. *Phys. Rev. Lett.* **42**, 129–132 (1979).
39. Lubin, P. M. & Smoot, G. F. Polarization of the cosmic background radiation. *Astrophys. J.* **245**, 1–17 (1981).
40. Lubin, P., Melese, P. & Smoot, G. Linear and circular polarization of the cosmic background radiation. *Astrophys. J.* **273**, L51–L54 (1983).
41. Sironi, G. *et al.* A 33 GHz polarimeter for observations of the cosmic microwave background. *New Astron.* **3**, 1–13 (1997).
42. Keating, B. G. *et al.* A limit on the large angular scale polarization of the cosmic microwave background. *Astrophys. J.* **560**, L1–L4 (2001).
43. Wollack, E. J., Jarosik, N. C., Netterfield, C. B., Page, L. A. & Wilkinson, D. A measurement of the anisotropy in the cosmic microwave background radiation at degree angular scales. *Astrophys. J.* **419**, L49–L52 (1993).
44. Hedman, M. M. *et al.* New limits on the polarized anisotropy of the cosmic microwave background at subdegree angular scales. *Astrophys. J.* **573**, L73–L76 (2002).
45. Cartwright, J. K. *et al.* Polarization observations with the cosmic background imager. In *Moriond Workshop 37, The Cosmological Model* (in the press).
46. de Oliveira-Costa, A. *et al.* First attempt at measuring the CMB cross-polarization. *Phys. Rev. D* (submitted); preprint astro-ph/0204021 at (<http://xxx.lanl.gov>) (2002).
47. Partridge, R. B., Richards, E. A., Fomalont, E. B., Kellermann, K. I. & Windhorst, R. A. Small-scale cosmic microwave background observations at 8.4 GHz. *Astrophys. J.* **483**, 38–50 (1997).
48. Subrahmanyam, R., Keesteven, M. J., Ekers, R. D., Sinclair, M. & Silk, J. An Australia telescope survey for CMB anisotropies. *Mon. Not. R. Astron. Soc.* **315**, 808–822 (2000).
49. Leitch, E. M. *et al.* Experiment design and first season observations with the degree angular scale interferometer. *Astrophys. J.* **568**, 28–37 (2002).
50. Pryke, C. *et al.* Cosmological parameter extraction from the first season of observations with the degree angular scale interferometer. *Astrophys. J.* **568**, 46–51 (2002).
51. Leitch, E. M. *et al.* Measurement of polarization with the Degree Angular Scale Interferometer. *Nature* **420**, 763–771 (2002).
52. Haslam, C. G. T. *et al.* A 408 MHz all-sky continuum survey. I—Observations at southern declinations and for the north polar region. *Astron. Astrophys.* **100**, 209–219 (1981).
53. Bond, J. R., Jaffe, A. H. & Knox, L. Estimating the power spectrum of the cosmic microwave background. *Phys. Rev. D* **57**, 2117–2137 (1998).
54. Seljak, U. & Zaldarriaga, M. A line-of-sight integration approach to cosmic microwave background anisotropies. *Astrophys. J.* **469**, 437–444 (1996).
55. White, M., Carlstrom, J. E., Dragovan, M. & Holzapfel, W. H. Interferometric observation of cosmic microwave background anisotropies. *Astrophys. J.* **514**, 12–24 (1999).
56. Hedman, M. M., Barkats, D., Gunderson, J. O., Staggs, S. T. & Winstein, B. A limit on the polarized anisotropy of the cosmic microwave background at subdegree angular scales. *Astrophys. J.* **548**, L111–L114 (2001).
57. Knox, L. Cosmic microwave background anisotropy window functions revisited. *Phys. Rev. D* **60**, 103516 (1999).
58. Halverson, N. W. *A Measurement of the Cosmic Microwave Background Angular Power Spectrum with DASI*. PhD thesis, Caltech (2002).
59. Tegmark, M. & de Oliveira-Costa, A. How to measure CMB polarization power spectra without losing information. *Phys. Rev. D* **64**, 063001 (2001).
60. Christensen, N., Meyer, R., Knox, L. & Luey, B. Bayesian methods for cosmological parameter estimation from cosmic microwave background measurements. *Class. Quant. Gravity* **18**, 2677–2688 (2001).
61. Wang, X., Tegmark, M. & Zaldarriaga, M. Is cosmology consistent? *Phys. Rev. D* **65**, 123001 (2002).
62. Wright, A. E., Griffith, M. R., Burke, B. F. & Ekers, R. D. The Parkes-MIT-NRAO (PMN) surveys. 2:

- Source catalog for the southern survey ($-87.5^\circ < \delta < -37^\circ$). *Astrophys. J. Suppl.* **91**, 111–308 (1994).
63. Mason, B. S., Pearson, T. J., Readhead, A. C. S., Shepherd, M. C. & Sievers, J. L. The anisotropy of the microwave background to $l = 3500$: Deep field observations with the cosmic background imager. *Astrophys. J.* (submitted); preprint astro-ph/0205384 at (<http://xxx.lanl.gov>) (2002).
 64. Zukowski, E. L. H., Kronberg, P. P., Forkert, T. & Wielebinski, R. Linear polarization measurements of extragalactic radio sources at $\lambda 6.3$ cm. *Astron. Astrophys. Suppl.* **135**, 571–577 (1999).
 65. Simard-Normandin, M., Kronberg, P. P. & Neidhoefer, J. Linear polarization observations of extragalactic radio sources at 2 cm and at 17–19 cm. *Astron. Astrophys. Suppl.* **43**, 19–22 (1981).
 66. Simard-Normandin, M., Kronberg, P. P. & Button, S. The Faraday rotation measures of extragalactic radio sources. *Astrophys. J. Suppl.* **45**, 97–111 (1981).
 67. Hildebrand, R. H. *et al.* A primer on far-infrared polarimetry. *Publ. Astron. Soc. Pacif.* **112**, 1215–1235 (2000).
 68. Lazarian, A. & Prunet, S. *AIP Conf. Proc.* **609**, *Astrophysical Polarized Backgrounds* (eds Cecchini, S., Cortiglioni, S., Sault, R. & Sbarra, C.) 32–43 (AIP, Melville, New York, 2002).
 69. Tegmark, M., Eisenstein, D. J., Hu, W. & de Oliveria-Costa, A. Foregrounds and forecasts for the cosmic microwave background. *Astrophys. J.* **530**, 133–165 (2000).
 70. Giardino, G. *et al.* Towards a model of full-sky Galactic synchrotron intensity and linear polarisation: A re-analysis of the Parkes data. *Astron. Astrophys.* **387**, 82–97 (2002).
 71. Gaensler, B. M. *et al.* Radio polarization from the inner galaxy at arcminute resolution. *Astrophys. J.* **549**, 959–978 (2001).
 72. Gray, A. D. *et al.* Radio polarimetric imaging of the interstellar medium: Magnetic field and diffuse ionized gas structure near the W3/W4/W5/HB 3 complex. *Astrophys. J.* **514**, 221–231 (1999).
 73. Platania, P. *et al.* A determination of the spectral index of galactic synchrotron emission in the 1–10 GHz range. *Astrophys. J.* **505**, 473–483 (1998).
 74. Banday, A. J. & Wolfendale, A. W. Fluctuations in the galactic synchrotron radiation—I. Implications for searches for fluctuations of cosmological origin. *Mon. Not. R. Astron. Soc.* **248**, 705–714 (1991).
 75. Zaldarriaga, M. The nature of the E-B decomposition of CMB polarization. *Phys. Rev. D* (submitted); preprint astro-ph/0106174 at (<http://xxx.lanl.gov>) (2001).

Acknowledgements We are grateful for the efforts of B. Reddall and E. Sandberg, who wintered over at the National Science Foundation (NSF) Amundsen–Scott South Pole research station to keep DASI running smoothly. We are indebted to M. Dragovan for his role in making DASI a reality, and to the Caltech CBI team led by T. Readhead, in particular, to S. Padin, J. Cartwright, M. Shepherd and J. Yamasaki for the development of key hardware and software. We are indebted to the Center for Astrophysical Research in Antarctica (CARA), in particular to the CARA polar operations staff. We are grateful for contributions from K. Coble, A. Day, G. Drag, J. Kooi, E. LaRue, M. Loh, R. Lowenstein, S. Meyer, N. Odalen, R. Pernic, D. Pernic and E. Pernic, R. Spotz and M. Whitehead. We thank Raytheon Polar Services and the US Antarctic Program for their support of the DASI project. We have benefitted from many interactions with the Center for Cosmological Physics members and visitors. In particular, we gratefully acknowledge many conversations with W. Hu on CMB polarization and suggestions from S. Meyer, L. Page, M. Turner and B. Winstein on the presentation of these results. We thank L. Knox and A. Kosowsky for bringing the Markov technique to our attention. We thank the observatory staff of the Australia Telescope Compact Array, in particular B. Sault and R. Subrahmanyan, for providing point source observations of the DASI fields. This research was initially supported by the NSF under a cooperative agreement with CARA, a NSF Science and Technology Center. It is currently supported by an NSF-OPP grant. J.E.C. gratefully acknowledges support from the James S. McDonnell Foundation and the David and Lucile Packard Foundation. J.E.C. and C.P. gratefully acknowledge support from the Center for Cosmological Physics.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.M.K. (e-mail: john@hyde.uchicago.edu).

Long-term *in vivo* imaging of experience-dependent synaptic plasticity in adult cortex

Joshua T. Trachtenberg^{*†}, Brian E. Chen^{*†}, Graham W. Knott[‡], Guoping Feng[§], Joshua R. Sanes[§], Egbert Welker[‡] & Karel Svoboda^{*}

^{*} Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA

[‡] Institut de Biologie Cellulaire et de Morphologie, Université de Lausanne, Rue du Bugnon 9, CH 1005, Lausanne, Switzerland

[§] Department of Anatomy and Neurobiology, Washington University School of Medicine, St Louis, Missouri 63110, USA

[†] These authors contributed equally to this work

Do new synapses form in the adult cortex to support experience-dependent plasticity? To address this question, we repeatedly imaged individual pyramidal neurons in the mouse barrel cortex over periods of weeks. We found that, although dendritic structure is stable, some spines appear and disappear. Spine lifetimes vary greatly: stable spines, about 50% of the population, persist for at least a month, whereas the remainder are present for a few days or less. Serial-section electron microscopy of imaged dendritic segments revealed retrospectively that spine sprouting and retraction are associated with synapse formation and elimination. Experience-dependent plasticity of cortical receptive fields was accompanied by increased synapse turnover. Our measurements suggest that sensory experience drives the formation and elimination of synapses and that these changes might underlie adaptive remodelling of neural circuits.

The cellular and synaptic mechanisms underlying experience-dependent plasticity in the adult neocortex are not understood^{1–7}. In the most widely studied cellular model of plasticity, namely use-dependent potentiation and depression of synaptic transmission, synaptic connectivity is assumed to be fixed^{8,9}. Another view holds that cortical plasticity results from a restructuring of neural circuits¹⁰; this is supported by studies documenting the growth of axonal and dendritic processes^{6,11,12}. A separate model suggests changes in cortical connectivity at the level of individual synapses, without large-scale rearrangements of neuronal processes^{13–15}. These models are probably not exclusive but might operate over distinct timescales¹⁶.

Here we focus on dendritic spines as the substrate of plasticity. Spines are tiny dendritic protrusions that receive the vast majority of excitatory synapses^{17,18}. In cultured preparations, spines grow in response to bursts of synaptic stimulation^{19,20} and make synapses²¹. In the adult brain, spines can appear in response to hormonal changes²² and prolonged sensory overstimulation¹⁵; during development, spines show experience-dependent structural plasticity²³. Furthermore, theoretical studies have suggested that the formation and elimination of synapses through growth of spines constitute a potential memory mechanism with enormous information capacity¹⁴.

Here we report that spines appear and disappear frequently in the adult cortex, whereas the growth or retraction of dendritic or axonal processes was not observed. Spine sprouting is associated with synapse formation, and spine retraction with synapse elimination. The rate of synaptic turnover is increased in response to novel sensory experience.

Imaging of dendritic spines in adult cortex

To determine whether synaptic connectivity in the adult cortex is fixed, or whether the formation and elimination of synapses occur, we developed a preparation for chronic high-resolution imaging of neuronal structure *in vivo*. We used mice expressing enhanced green fluorescent protein in a small subpopulation (~1%) of layer-5 pyramidal neurons²⁴. Young adult mice (6–10 weeks of age) were prepared for imaging by implanting a small imaging window

centred over the barrel cortex (see Methods) (Fig. 1a). Two-photon laser-scanning microscopy^{23,25,26} revealed fluorescent dendritic arbors, in a pattern of labelling that resembles a Golgi stain (Fig. 1b–e). Characteristically for layer-5 neurons, these arbors were in layers 1 and 2 (ref. 27). At higher magnifications, dendritic spines (Fig. 1f), axons and presynaptic terminals (see section ‘New spines form synapses’) could be imaged with high contrast. In a typical long-term experiment the same dendritic regions were imaged every day for 8 days and less frequently thereafter (Fig. 1f).

Dynamic spines in the adult cortex

Previous time-lapse imaging studies in the developing barrel cortex demonstrated that spines show rapid motility, and grow and retract over times of tens of minutes²³. Relatively few changes were found over these periods in the adult cortex (see Methods). However, chronic time-lapse imaging revealed that spines were still dynamic, but over much longer timescales (see Fig. 1f). About 20% of spines disappeared between imaging sessions from one day to the next, balanced by the formation of new spines. About 60% of spines persisted for at least 8 days. We examined whether growth or retraction also occurred at the level of dendritic branching. Of 125 dendritic branches imaged at high magnification over timescales as long as 1 month (range 8–32 days), we failed to observe any branch additions or subtractions (Fig. 2). Similar conclusions hold for axonal branches that were sometimes apparent in our images (data not shown). Although we cannot rule out the possibility of very slow (less than μm per day) growth of dendritic and axonal branch tips, large-scale changes in the arrangement of neuronal processes do not occur and changes in dendritic structure are limited mainly to spines.

New spines form synapses

Does the growth and retraction of spines correspond to synaptogenesis and synapse elimination? To address this question we combined chronic *in vivo* imaging with ultrastructural analysis. Two dendritic regions that had been previously imaged for eight sequential days *in vivo* were identified in fixed tissue (Fig. 3A–D) and reconstructed using serial-section electron microscopy (SSEM);

Supplementary Information). The general appearance of these reconstructed dendrites (Fig. 3E), characterized by a relatively high proportion of shaft synapses and low spine densities, is consistent with previous work on layer 1 (ref. 28). One region of dendrite (length 29 μm) contained 22 synapses (18 asymmetric, 4 symmetric) (Fig. 3E) and two spines that had been observed to appear between the last two imaging sessions (days 7 and 8) (boxed

regions in Fig. 3D and E). The second region (not shown; length 21 μm) contained 17 synapses (14 asymmetric and 3 symmetric) and two spines that appeared between the last two imaging sessions. In total, 18 spines were reconstructed, 4 of which were new spines (less than 1 day old). Comparing SSEM reconstructions with the *in vivo* images showed that two new spines contained clear synapses (Fig. 3F and G). These spines contacted presynaptic boutons and

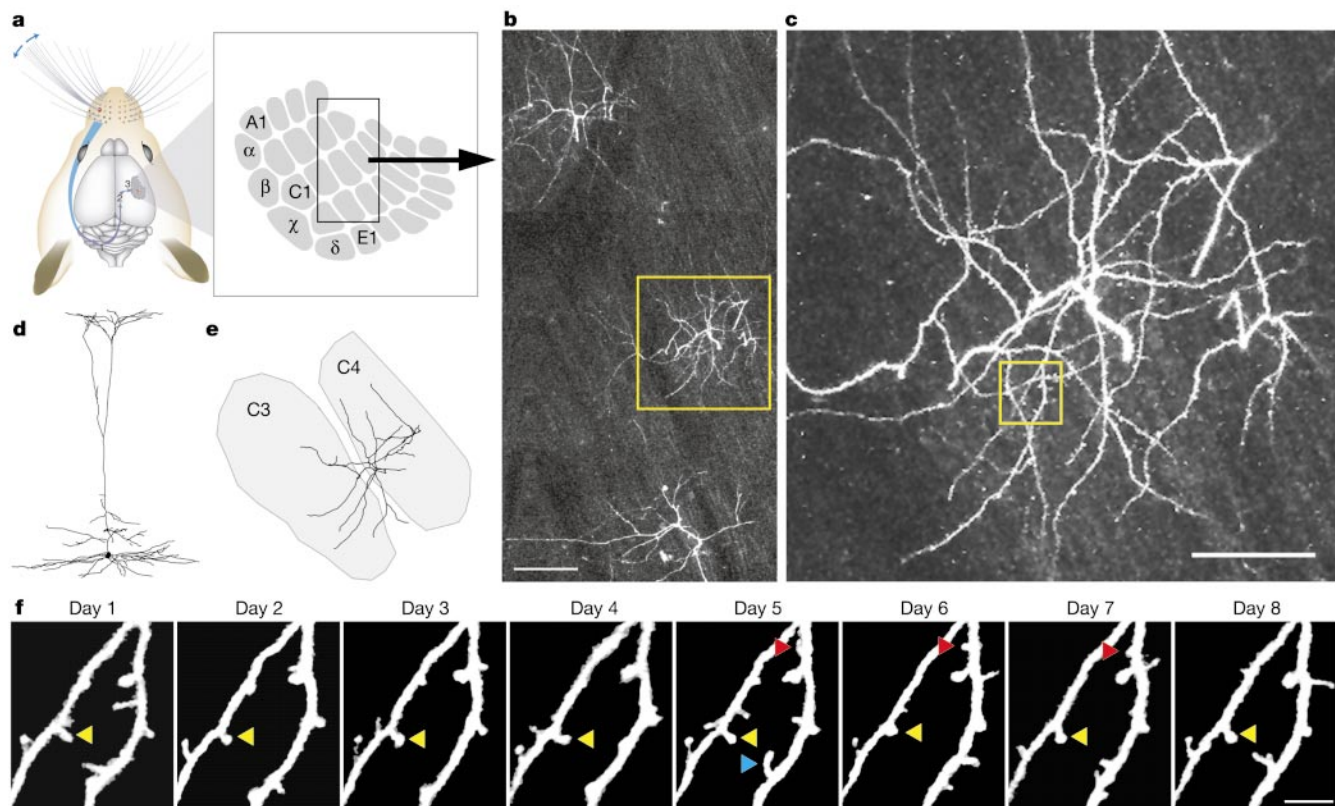


Figure 1 Chronic time-lapse imaging of dendritic spines in the barrel cortex *in vivo*. **a**, Diagram of the barrel cortex and its topographic structure (boxed region, right). **b**, Low-magnification dorsal view of the apical dendrites of three layer-5 pyramidal neurons. Scale bar, 100 μm . **c**, Higher-magnification view of one apical tuft (yellow box in **b**). Scale bar, 50 μm . **d**, Neuron reconstructed from images of fixed sections obtained after collection of

the last vital image (same neuron as in **c**). **e**, Dorsal view of the reconstructed apical tuft, depicting the positions of each dendritic branch relative to the barrel map. **f**, High-resolution time-lapse images of a dendritic region (yellow box in **c**). Examples of transient, semi-stable and stable spines are labelled with blue, red and yellow arrowheads, respectively. Scale bar, 5 μm .

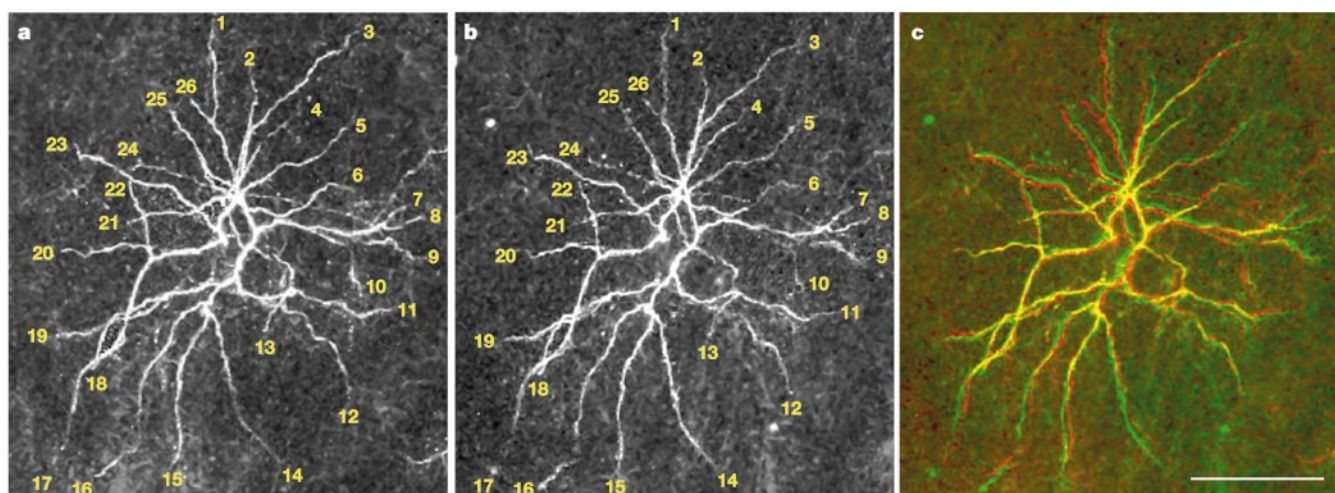


Figure 2 Dendritic branches are stable over weeks. **a**, Dorsal view of an apical tuft at imaging day 16; 26 branch tips are labelled for reference. **b**, Imaging day 32. **c**, Overlay (day 16, red; day 32, green). Scale bar, 100 μm .

were apposed to active zones containing clusters of synaptic vesicles (Fig. 3H and I). Thus, new spines in the adult brain make synapses.

Two views of the relationship between spine turnover and synapses are currently prominent. In the first, spines could grow

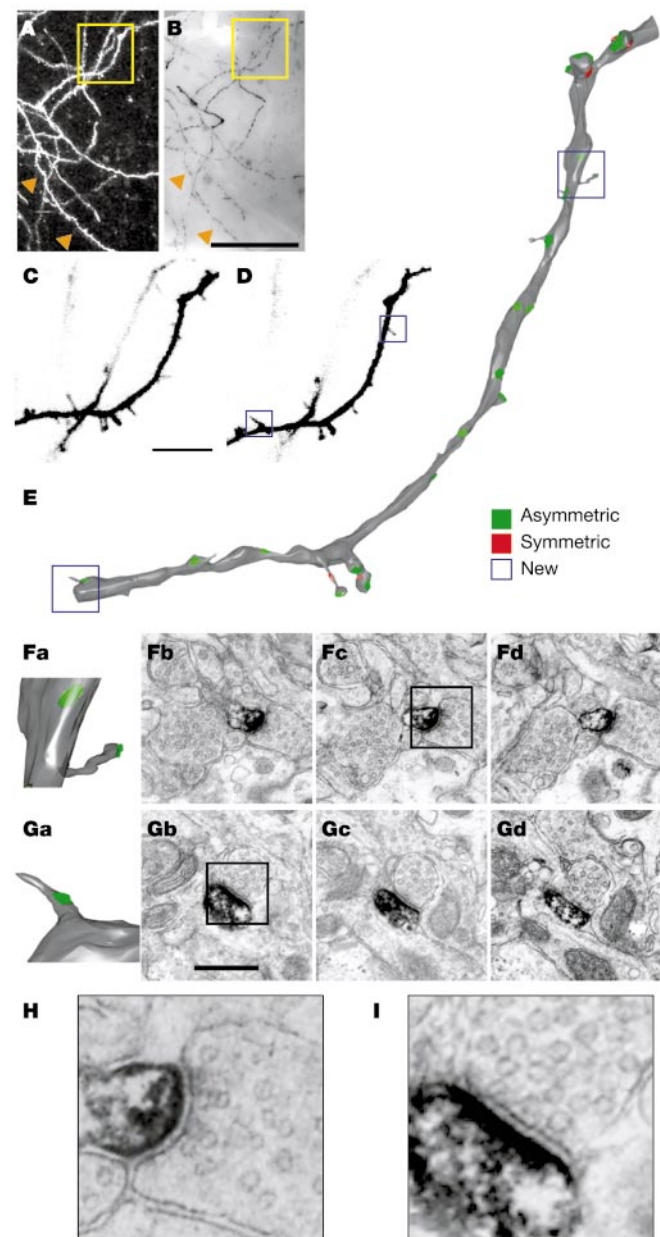


Figure 3 New spines establish synapses. **A**, *In vivo* image of a dendritic region on the eighth and final imaging day. **B**, The same dendritic region immunolabelled for green fluorescent protein in fixed tissue. Orange arrowheads identify the same dendrite in the two panels. The yellow box indicates the dendritic region that was serially reconstructed for electron microscopy. Scale bar, 50 μm . **C**, *In vivo* image (inverted), acquired on the seventh imaging day (yellow box in **A** and **B**). **D**, Same dendritic region as in **C** on the eighth imaging day, immediately before perfusion. Two new spines appeared after the seventh imaging day (boxes). Scale bar, 10 μm . **E**, Three-dimensional reconstruction of the dendritic region from serial-section electron micrographs. The red regions indicate symmetric (presumably inhibitory) synapses; the green regions asymmetric (presumably excitatory) synapses. The boxes delineate the same two new spines as in **D**. **Fa**, **Ga**, New spines identified by the upper-right and lower-left boxes in **E**, respectively. Green regions identify asymmetric synapses. **Fb-d**, **Gb-d**, Electron micrographs showing synapses established by these new spines. Scale bar, 0.5 μm . **H**, **I**, Enlargement of boxed regions of **Fc** and **Gb**, respectively, showing synaptic clefts and vesicle clusters. Scale bar, 0.125 μm .

to make new synapses^{13,14}. In the second, spines could emerge just below an existing shaft synapse and drag this synapse along, in which case spine growth would not be associated with a gain of synapses¹⁷.

Three lines of evidence support the first model. First, in two examples from two separate animals we observed spine growth towards axonal varicosities (Fig. 4). In these examples the varicosities were stable and were contacted, at the level of light microscopy, on two occasions by different spines emerging from the same dendrite. Second, if the formation of spines were to push synapses away from the dendritic shaft, local movements and changes in the curvature of axons would be observable. However, the trajectory of axons was remarkably constant (Fig. 4, and data not shown).

The third line of evidence comes from comparing the number of spines that retracted over 8 days of imaging with the number of shaft synapses observed in the SSEM reconstructions. If retracting spines did indeed pull their associated presynaptic bouton to establish shaft synapses, we would expect the number of shaft synapses to be greater than or equal to the number of retracted spines. We counted 17 excitatory shaft synapses in the two dendritic segments reconstructed by SSEM. In contrast, in these same dendritic regions we observed 28 spines that retracted over 8 days of imaging; the actual number of retracted spines is likely to be considerably larger because we cannot reliably resolve, and hence did not analyse, spines that project above and below the dendrite. Thus, the number of retracted spines is severalfold higher than the number of shaft synapses. Furthermore, of the retracted spines only seven were within 0.5 μm of a shaft synapse. We conclude that retracting spines do not maintain synaptic contact. In addition, two spines that appeared between the last two imaging sessions seemed to be in various phases of synapse formation or elimination. One new spine clearly contacted an axon, although it lacked presynaptic specialization, and one spine did not show a clear indication of presynaptic contact in the SSEM sections (data not shown). Thus, *in vivo* imaging in combination with SSEM provides strong evidence for *de novo* synaptogenesis and synapse elimination in the adult neocortex.

Spines are distributed into kinetic classes

Our time-lapse images indicate that stable and dynamic spines are apparently intermixed along the dendrite (Figs 1f and 5a). Despite spine turnover, spine density was stable (Fig. 5b). To quantify the

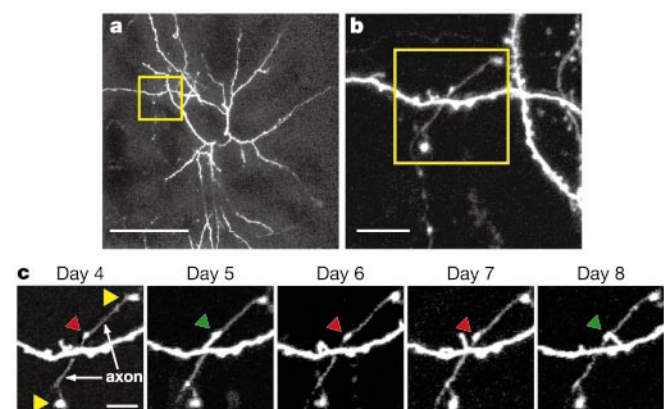


Figure 4 *In vivo* imaging of putative synapse formation and elimination. **a**, Low-magnification image. Scale bar, 100 μm . **b**, Higher-magnification view (yellow box in **a**). Scale bar, 10 μm . **c**, Further magnified view (yellow box in **b**). An axon (day 4, white arrows) with varicosities (day 4, red and yellow arrowheads) is coursing across the dendrite. One varicosity (day 4, red arrowhead) seems to be contacted by a dendritic spine on two separate occasions (days 5 and 8). The transition of the arrowhead from red to green indicates the close apposition of the axon and two different dendritic spines, suggesting synapse formation. Scale bar, 5 μm .

kinetics of spine turnover we measured the lifetime of each spine as the number of sequential days over which a spine was present during an 8-day period (1224 spines, $n = 7$ cells, one cell per animal). The distribution of spine lifetimes revealed at least three kinetic components (Fig. 5c): transient spines appeared in only one image (lifetime ≤ 1 day); semi-stable spines persisted with a mean lifetime of 2–3 days (Fig. 5c; exponential fit to the central portion of the lifetime distribution); and stable spines survived the entire imaging period (lifetime ≥ 8 days). On any given day, transient spines make up $\sim 17\%$ of the spine population, semi-stable spines $\sim 23\%$ and stable spines 60% .

We examined whether stable spines are permanent or whether they, too, turn over with time. In three animals we imaged spines for eight sequential days and identified stable spines. We then imaged these same spines 20–24 days later (~ 30 days after the first imaging day). Of 157 spines that were present throughout the first eight imaging days, 132 were still present on day 30 (cf. Figure 5d, yellow arrowheads). A substantial fraction ($\sim 15\%$) of stable spines had therefore disappeared (cf. Figure 5d, orange arrowhead). An analysis of spine structure revealed that large spines (as judged by volume²⁹) were more likely to be stable than transient spines (Supplementary Information). Large mushroom-shaped spines³⁰ were almost always stable for the entire duration of the experiment (for example, yellow arrowheads in Figs 1f and 5a, d).

Sensory experience modulates spine turnover

Are the addition and subtraction of spines influenced by sensory experience? To explore this question we altered sensory experience by trimming every other whisker on the mystacial pad, such that each trimmed whisker was surrounded by intact whiskers (Fig. 6a). This ‘chessboard’ deprivation produces an imbalance in the activation of neighbouring cortical columns, similar to the monocular deprivation used in the visual system³¹, and induces a rapid and robust remodelling of whisker representation^{32,33}.

We tested whether the dynamics of spines is influenced by chessboard deprivation. In each mouse, whiskers were left intact for the first 4 days of imaging. Whiskers on the contralateral mystacial pad were then trimmed, and imaging continued for a further 4 days (Fig. 6a). A comparison of spine lifetime in the barrel cortex before and after deprivation revealed considerable experience-dependent effects (579 spines, $n = 4$ cells). Chessboard deprivation increased the pool of transient spines, present for only a single day or less, with a concomitant decrease in the pool of stable spines (Fig. 6b). This change was specific to deprived cortical areas, because no change in spine lifetimes was observed in cells outside the barrel cortex (645 spines, $n = 3$ cells; Fig. 6c). Sensory deprivation did not modulate spine density regardless of whether a cell was in the barrel cortex (Fig. 6d). The lifetime of spines is therefore experientially modulated and homeostatic mechanisms keep spine

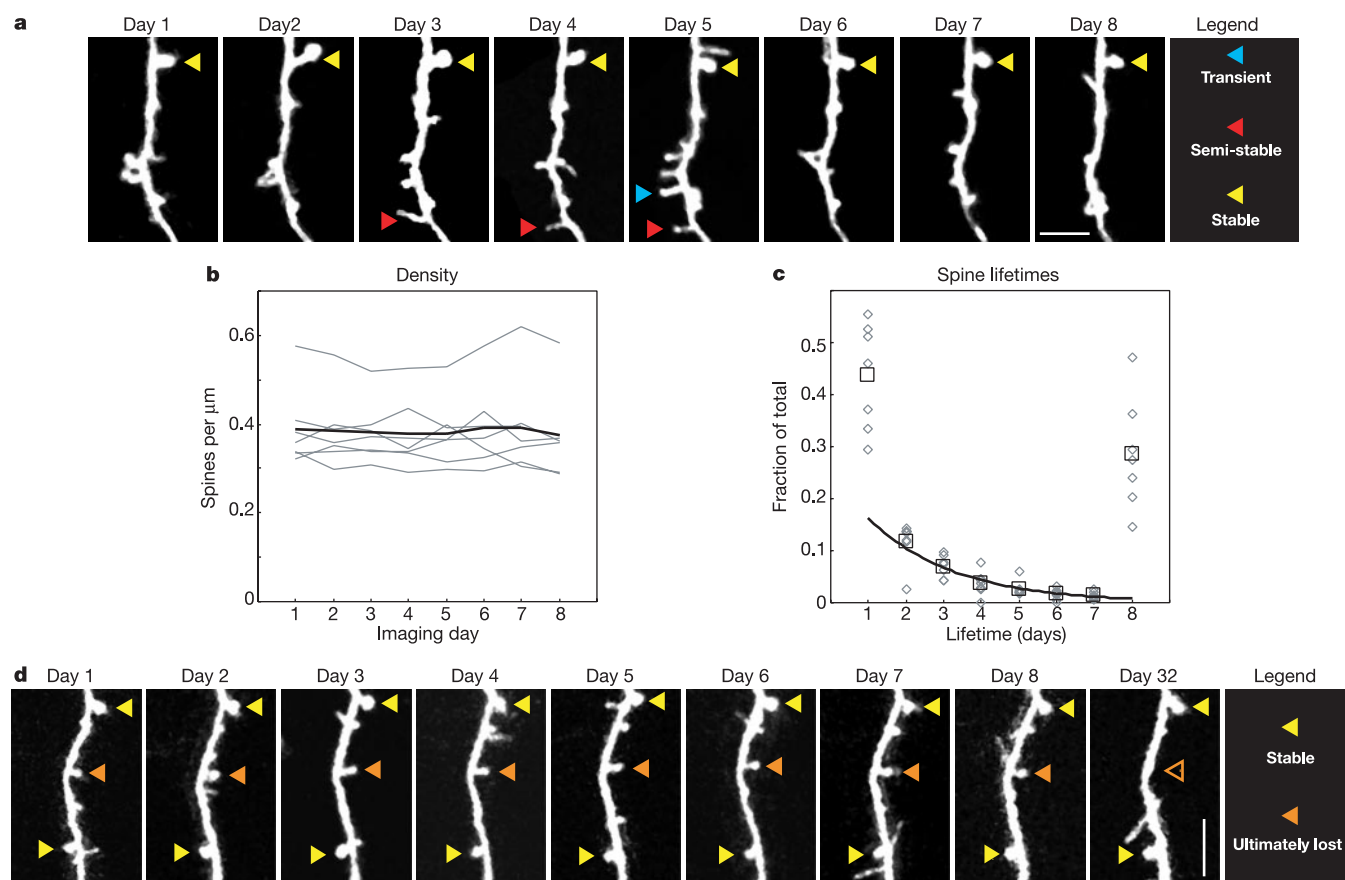


Figure 5 Spines appear and disappear with broadly distributed lifetimes, without changing spine density. **a**, Images of a dendritic segment acquired over eight sequential days. Examples of transient, semi-stable and stable spines (with lifetimes of ≤ 1 day, 2–7 days, and ≥ 8 days, respectively) are indicated with blue, red and yellow arrowheads, respectively. Scale bar, $5 \mu\text{m}$. **b**, Spine density (spines per $100 \mu\text{m}$ of dendritic length) is plotted for seven neurons (grey lines) for each imaging day; the average spine density is indicated by a black line. **c**, Distribution of spine lifetimes. Lifetimes are defined as the number of sequential days (from a total of eight) over which a spine existed. Individual

neurons (grey diamonds) and the average (black squares) are shown. The fraction of spines with lifetimes of 2–7 days is fitted with a single time constant (thick black line). The fractions of spines with lifetimes of less than 1 day (transient spines) and greater than 8 days (stable spines) are significantly greater than predicted from the exponential fit, and therefore constitute distinct kinetic populations. **d**, Example of a stable spine that ultimately disappears. Two large mushroom spines, indicated by the yellow arrowheads, have lifetimes of ≥ 32 days, whereas a third spine, indicated by the orange arrowhead, persists for ≥ 8 days but disappears by day 32. Scale bar, $5 \mu\text{m}$.

and synaptic densities constant.

To examine the time course of the onset of synaptic rearrangement in these cells, we measured the fraction of spines gained and lost from one day to the next (turnover ratio) throughout the control and experimental conditions (Fig. 6e). The turnover ratio was stable over the first four imaging days in which whiskers were intact. Chessboard deprivation for 24 h failed to induce a significant change in the fraction of spines gained or lost relative to the control period. However, 2–4 days after the onset of chessboard deprivation, significant increases occurred in spine turnover relative to both the control period and to the first 24 h after deprivation ($P < 0.05$ for each comparison; two-tailed t -test corrected for multiple comparisons). Consistent with input-specific plasticity was the observation that spine turnover was stable in cells lying outside the deprived barrel cortex (Fig. 6e).

Plasticity of receptive fields

To examine how synaptic circuits are modified in response to chessboard deprivation that induces changes in spine turnover (Fig. 6), we used membrane potential measurements^{34–36} (Supplementary Information). Intracellular recordings were made from pyramidal neurons (four control mice and nine mice that had

experienced 3–4 days of chessboard deprivation). To determine the neuron's position in the barrel map and therefore its principal whisker, recorded neurons were subsequently identified in cytochrome-oxidase-stained sections and assigned to a barrel column³⁴. The amplitudes of postsynaptic potentials evoked by deflections of individual whiskers were used to characterize receptive field organization (Fig. 7a). In control animals, the structure of receptive fields resembled those of receptive fields previously measured in rats (Fig. 7b)^{34–36}. Neurons responded most strongly to the principal whisker and more weakly to stimulation of first-order and second-order surround whiskers. In chessboard-deprived animals, neurons in barrels with their principal whisker trimmed had a significantly enhanced surround receptive field, corresponding to spared whiskers (Wilcoxon signed-rank test, $P < 0.05$). These measurements show that experience-dependent reorganization of synaptic circuits underlies receptive field plasticity. This is consistent with a role for synapse formation and elimination as the cellular mechanism underlying experience-dependent plasticity.

Discussion

A central question in neuroscience is how and to what extent neurons and their synapses change to support experience-dependent functional changes in the adult neocortex. To address this issue we used two-photon laser-scanning microscopy for chronic imaging of layer-5 pyramidal cell axons, dendrites and their spines in the barrel cortex *in vivo* (Fig. 1). We find that spines appear and disappear over periods of days to weeks, even though their average density remains constant (Fig. 5). In contrast, the large-scale branching pattern of dendrites and axons was stable (Fig. 2). To determine whether the sprouting and retraction of spines contribute to the rearrangement of cortical synaptic circuits, we directly correlated time-lapse images of dendritic spines *in vivo* with SSEM reconstructions of the same structures (Fig. 3). These measurements demonstrate that spine sprouting is associated with synapse formation, and spine retraction with synapse elimination. Synaptogenesis is therefore not limited to development but also occurs at

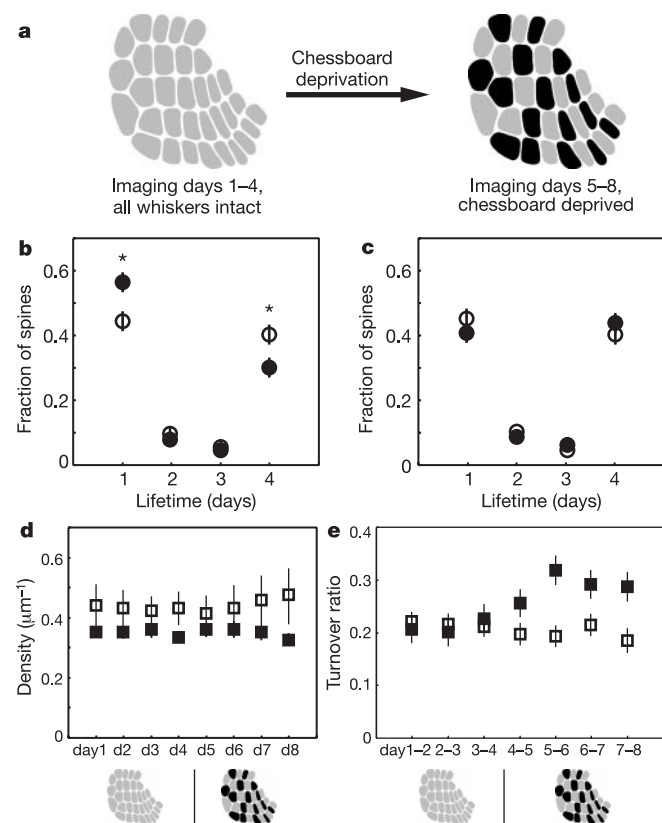


Figure 6 Altering sensory experience increases spine turnover rates. **a**, Experimental protocol (see text). Black barrels are deprived. **b**, Distribution of spine lifetimes (over four imaging days) for control (open circles) and chessboard-deprived (solid circles) conditions for cells within the barrel cortex. Asterisks indicate significantly different populations. Deprivation reduces the fraction of spines with longer lifetimes and increases the number of transient spines. **c**, Spine lifetimes for neurons outside the barrel cortex. Deprivation does not change spine lifetimes. Symbols as in **b**. **d**, Spine density for cells lying within (solid squares) or outside (open squares) the barrel cortex. Spine density does not change in response to deprivation. **e**, Turnover ratio (the fraction of spines that turn over between successive imaging sessions) as a function of time. Chessboard deprivation occurred immediately after imaging day 4. Turnover ratio increased after deprivation within (solid squares) but not outside (open squares) the barrel cortex. Error bars in **b–e** show s.e.m.

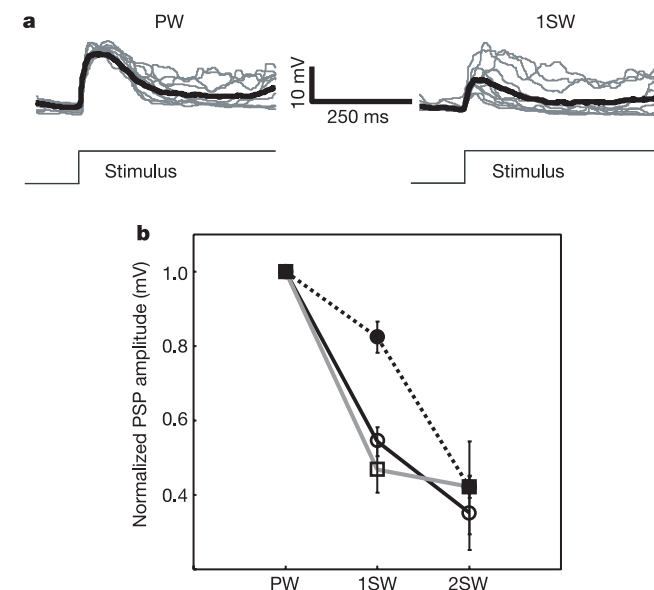


Figure 7 Experience-dependent receptive field plasticity. **a**, Examples of postsynaptic potentials (grey lines) and averages (thick black lines) evoked by stimulation of principal whisker (PW) and first-order surround whisker (1SW). **b**, Normalized evoked postsynaptic potential (PSP) amplitudes (means $\pm$ s.e.m.) for PW and surround whiskers (1SW, 2SW, first-order and second-order surround whiskers, respectively) for cells in control (black line), spared (grey line) and deprived (dotted line) barrels.

high rates in the adult brain. Reports of stable synaptic densities in the adult neocortex should not be viewed as the absence of synaptogenesis but rather as balanced rates of synapse formation and elimination^{37,38}.

We find that synaptic lifetimes are distributed broadly (Fig. 5). About 80% of synapses were detectable for a day or longer; about 60% belonged to the stable pool imaged for at least 8 days. Even this stable pool was found to turn over, with only ~50% of spines surviving for 30 days or longer. Assuming stochastic behaviour, we estimate that the mean lifetime of the stable pool would be on the order of 120 days. However, even longer imaging experiments might reveal that a subpopulation of synapses persists for the entire lifetime of the animal. Furthermore, our data indicate that synaptic stability is correlated with spine structure: larger spines tended to be more stable, and large mushroom spines were particularly so (Supplementary Information). Future experiments combining *in vivo* imaging with SSEM might reveal the ultrastructural determinants of synapse stability.

Our data are consistent with a model in which adaptive change in the adult neocortex occurs through changes in connectivity patterns through the sprouting and retraction of dendritic spines, without large-scale rearrangements of neuronal processes. Similar changes have been observed previously in developing brain slices^{19–21} and in the developing barrel cortex²³, as well as in response to long-term overstimulation of whiskers in the adult cortex¹⁵. New synapses could strengthen previously existing connections or connect previously unconnected neurons.

To determine whether synapse addition and subtraction might be involved in experience-dependent plasticity, we correlated receptive field changes with chronic imaging of dendritic spines. We find that after 2–4 days of sensory deprivation the turnover of spines increases markedly (Fig. 6), coincident with rearrangements of receptive field structure as probed by maps of excitatory synaptic potentials (Fig. 7). Taken together, our data suggest that changes in synaptic connectivity might underlie experience-dependent rewiring of the adult brain.

What is the role of transient synapses in experience-dependent rearrangements of cortical circuits? Our data are consistent with the following model. Spines sample (perhaps randomly) presynaptic partners from the set of available presynaptic partners (potential synapses)¹⁴. Appropriate synapses might then be stabilized in an activity-dependent manner, and other synapses are eliminated. In this model the transient increase in synapse turnover after the onset of deprivation is the consequence of destabilization of previously stable synapses.

Our results have relevance to understanding the mechanisms underlying the closure of critical periods. Critical periods, first characterized in the primary visual cortex³⁹, are temporal windows in postnatal development when experience profoundly influences the final organization of neural connections. Although we report a high rate of synaptogenesis and synapse elimination in the adult cortex that is regulated by experience, all of these synaptic changes are local, without growth or elimination of axons or dendritic branches, using a fixed complement of potential synapses. This is in profound contrast to critical period plasticity, which has been documented to be associated with changes at the level of axonal⁴⁰ and dendritic branching (M. Maravall & K.S., unpublished observations). The cessation of such process growth might bring critical periods to a close. This would ensure that plasticity induced in the adult cortex would be reversible, because the original presynaptic partners remain in place, as do the major dendritic branches. In contrast, plasticity induced and maintained throughout the duration of a critical period would be irreversible in the adult because axonal and dendritic branch growth and retraction would fundamentally restructure the circuit. □

Methods

Surgery

Male c57/Bl6 transgenic mice (postnatal day (PND) 34–74) expressing enhanced green fluorescent protein in a subset of cortical-layer-5 pyramidal neurons (transgenic line M; ref. 24) were used for this study. Mice were deeply anaesthetized with an intraperitoneal injection of ketamine (0.13 mg per g body weight) and xylazine (0.01 mg g⁻¹). Dexamethasone (0.02 ml at 4 mg ml⁻¹) was administered subcutaneously. A 5 × 5 mm² region of the skull was removed, exposing the dura. An optical chamber was then constructed by covering the intact dura with agarose (Sigma, catalogue no. A9793) and a custom-made cover glass (no. 1), sealed with dental acrylic. A small titanium bar with tapped screw holes was embedded into the acrylic to stabilize the animal for subsequent imaging sessions. Animals were returned to their cages after surgery and allowed 7–10 days to recover before daily imaging began.

Imaging

In vivo images of neurons expressing green fluorescent protein were acquired as described²³ by using a 40 × 0.8 numerical aperture objective lens (Zeiss) and custom image-acquisition software (MatLab). For chronic experiments, animals were lightly anaesthetized with ketamine/xylazine at half the surgical dose (see above). At this level, anaesthesia typically wore off within 30 min, at which time animals were again returned to their cage. In acute experiments, performed in anaesthetized animals over durations of hours, we saw few changes (Supplementary Information). The apical dendrites of single layer-5 pyramids (three to eight fields, 45 × 45 μm²) were imaged daily over periods of 8–10 days and less frequently thereafter. Care was taken with each imaging session to achieve similar fluorescence levels. Imaged dendrites were within 100 μm of the pia and were therefore in layer 1. Standard histological techniques were used to determine whether individual neurons were in the barrel cortex³⁴.

Image analysis was done blind with regard to experimental condition. For every image stack a single montage image was created, composed of the best focal plane for each spine. The montage images were then aligned by using fiducial marks that were stable across all imaging days. The presence or absence of each spine was recorded each day and spine positions along the dendrites were measured. Turnover ratios were calculated as $(N_{\text{gained}} + N_{\text{lost}})/(2N_{\text{total}})$, where N_{gained} is the number of new spines, N_{lost} is the number of lost spines, and N_{total} is the total number of spines. Turnover ratios were indistinguishable in animals in which experiments were initiated at PND 34 or 74, indicating that, in terms of spine dynamics, our measurements report the adult state.

To determine whether spine densities (Figs 5b and 6d) changed over time, linear regression lines were fitted to the density plots for each cell. The significance of the difference between means of daily turnover rates (Fig. 6e) was determined with a Student's two-tailed *t*-test. To correct for multiple comparisons the reported *P* value was calculated as $1 - (1 - P)^n$, where *n* is the number of comparisons. To determine whether spine lifetimes were significantly different before and after chessboard deprivation, we calculated the predicted standard deviation in spine lifetimes from the binomial distribution as $\sqrt{[nP(1 - P)]}$. Significance was set at *P* = 0.05.

The axial resolution of optical microscopy is relatively poor⁴¹. We have therefore not analysed spines that projected mainly along the optical axis. Spines vary greatly in volume⁴² and therefore fluorescence intensity (~50-fold; Supplementary Information), the smallest spines being close to the detection limit. However, we are confident that we reliably detected even the smallest protrusions in our *in vivo* imaging. First, the spine density in our *in vivo* images (Fig. 6d) is not different from measurements in fixed tissue from the same animals ($0.48 \pm 0.16 \mu\text{m}^{-1}$). Second, even the smallest structures reconstructed in SSEM were also detected in our *in vivo* images (Fig. 3F and G). The converse was also true: spines that disappeared in time-lapse imaging were confirmed as absent in three-dimensional SSEM reconstructions (*n* = 6 spines). It is therefore unlikely that an apparent loss of spines corresponds to changes in the imaging conditions or to local torsions of the dendritic shaft.

To check for possible effects of the cranial window on the health of cortical circuits we conducted several control experiments. We performed measurements of membrane potential directly below the window, in animals that had had windows implanted 2 weeks before recording, and compared these measurements with those from age-matched controls (see below). As a measure of the health of cortical circuitry we recorded spontaneous synaptic activity and characterized the amplitudes of spontaneous fluctuations in membrane potential^{23,34}. We found that the amplitudes of membrane fluctuations (characterized as the s.d.) were identical in implanted animals (4.6 ± 2.6 mV; *n* = 9) and controls (5.0 ± 1.7 mV; *n* = 9), suggesting that our implants did not perturb the underlying cortical circuits. Consistently, densities of spines measured in fixed brains were not different from densities immediately below the imaging window and spine densities did not vary with experimental duration (Figs 5b and 6d). Finally, SSEM analysis of layer 1 did not show evidence for the proliferation of glial cells or other indications of tissue damage (Supplementary Information).

Electron microscopy and electrophysiology

To study the ultrastructure of the imaged dendrites, a pre-embedding immunolabelling protocol was used⁴³. Identified cells were drawn with Neurolucida software (MicroBrightfield) and the relevant dendrites were serially sectioned and photographed under the electron microscope for three-dimensional reconstructions (Supplementary Information). Measurements of subthreshold receptive fields were performed on wild-type c57/bl6 5-week-old male mice (*n* = 13) essentially as described³⁴ (Supplementary Information).

Received 18 July; accepted 23 October 2002; doi:10.1038/nature01273.

1. Squire, L. R. & Alvarez, P. Retrograde amnesia and memory consolidation: a neurobiological

- perspective. *Curr. Opin. Neurobiol.* **5**, 169–177 (1995).
2. Jenkins, W. M., Merzenich, M. M., Ochs, M. T., Allard, T. & Guic-Robles, E. Functional reorganization of primary somatosensory cortex in adult owl monkeys after behaviorally controlled tactile stimulation. *J. Neurophysiol.* **63**, 82–104 (1990).
3. Bakin, J. S. & Weinberger, N. M. Classical conditioning induces CS-specific receptive field plasticity in the auditory cortex of the guinea pig. *Brain. Res.* **536**, 271–286 (1990).
4. Gilbert, C. D. & Wiesel, T. N. Receptive field dynamics in adult primary visual cortex. *Nature* **356**, 150–152 (1992).
5. Diamond, M. E., Huang, W. & Ebner, F. F. Laminar comparison of somatosensory cortical plasticity. *Science* **265**, 1885–1888 (1994).
6. Florence, S. L., Taub, H. B. & Kaas, J. H. Large-scale sprouting of cortical connections after peripheral injury in adult macaque monkeys. *Science* **282**, 1117–1121 (1998).
7. Wang, X., Merzenich, M. M., Sameshima, K. & Jenkins, W. M. Remodelling of hand representation in adult cortex determined by timing of tactile stimulation. *Nature* **378**, 71–75 (1995).
8. Tanzi, E. I fatti e le induzioni nell'odierna istologia del sistema nervoso. *Riv. Sper. Freniatr.* **19**, 419–472 (1893).
9. Martin, S. J., Grimwood, P. D. & Morris, R. G. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annu. Rev. Neurosci.* **23**, 649–711 (2000).
10. Ramon y Cajal, S. Neue Darstellung vom histologischen Bau des Centralnervensystems. *Arch. Anat. Physiol. Anat. Abt. Suppl.*, 319–428 (1893).
11. Darian-Smith, C. & Gilbert, C. D. Axonal sprouting accompanies functional reorganization in adult cat striate cortex. *Nature* **368**, 737–740 (1994).
12. Volkmar, F. R. & Greenough, W. T. Differential rearing effects on rat visual cortical plasticity. *Science* **176**, 1445–1447 (1972).
13. Ziv, N. E. & Smith, S. J. Evidence for a role of dendritic filopodia in synaptogenesis and spine formation. *Neuron* **17**, 91–102 (1996).
14. Stepanyants, A., Hof, P. R. & Chklovskii, D. B. Geometry and structural plasticity of synaptic connectivity. *Neuron* **34**, 275–288 (2002).
15. Knott, G. W., Quairiaux, C., Genoud, C. & Welker, E. Formation of dendritic spines with GABAergic synapses induced by whisker stimulation in adult mice. *Neuron* **34**, 265–273 (2002).
16. Bailey, C. H. & Kandel, E. R. Structural changes accompanying memory formation. *Annu. Rev. Physiol.* **55**, 397–426 (1993).
17. Harris, K. M. Structure, development, and plasticity of dendritic spines. *Curr. Opin. Neurobiol.* **9**, 343–348 (1999).
18. Nimchinsky, E. A., Sabatini, B. L. & Svoboda, K. Structure and function of dendritic spines. *Annu. Rev. Physiol.* **64**, 313–353 (2002).
19. Maletic-Savatic, M., Malinow, R. & Svoboda, K. Rapid dendritic morphogenesis in CA1 hippocampal dendrites induced by synaptic activity. *Science* **283**, 1923–1927 (1999).
20. Engert, F. & Bonhoeffer, T. Dendritic spine changes associated with hippocampal long-term synaptic plasticity. *Nature* **399**, 66–70 (1999).
21. Toni, N., Buchs, P. A., Nikonenko, I., Bron, C. R. & Müller, D. LTP promotes formation of multiple spine synapses between a single axon terminal and a dendrite. *Nature* **402**, 421–425 (1999).
22. Yankova, M., Hart, S. A. & Woolley, C. S. Estrogen increases synaptic connectivity between single presynaptic inputs and multiple postsynaptic CA1 pyramidal cells: A serial electron-microscopic study. *Proc. Natl Acad. Sci. USA* **98**, 3525–3530 (2001).
23. Lendvai, B., Stern, E., Chen, B. & Svoboda, K. Experience-dependent plasticity of dendritic spines in the developing rat barrel cortex *in vivo*. *Nature* **404**, 876–881 (2000).
24. Feng, G. *et al.* Imaging neuronal subsets in transgenic mice expressing multiple spectral variants of GFP. *Neuron* **28**, 41–51 (2000).
25. Denk, W., Strickler, J. H. & Webb, W. W. Two-photon laser scanning microscopy. *Science* **248**, 73–76 (1990).
26. Denk, W. & Svoboda, K. Photon upmanship: why multiphoton imaging is more than a gimmick. *Neuron* **18**, 351–357 (1997).
27. Kim, H. G. & Connors, B. W. Apical dendrites of the neocortex: correlation between sodium- and calcium-dependent spiking and pyramidal cell morphology. *J. Neurosci.* **13**, 5301–5311 (1993).
28. Vaughn, J. E. & Peters, A. A three dimensional study of layer I of the rat parietal cortex. *J. Comp. Neurol.* **149**, 355–370 (1973).
29. Svoboda, K., Tank, D. W. & Denk, W. Direct measurement of coupling between dendritic spines and shafts. *Science* **272**, 716–719 (1996).
30. Harris, K. M., Jensen, F. E. & Tsao, B. Three-dimensional structure of dendritic spines and synapses in rat hippocampus (CA1) at postnatal day 15 and adult ages: implications for the maturation of synaptic physiology and long-term potentiation. *J. Neurosci.* **12**, 2685–2705 (1992).
31. Wiesel, T. N. The postnatal development of the visual cortex and the influence of environment. *Nature* **299**, 583–591 (1982).
32. Glazewski, S., McKenna, M., Jacquin, M. & Fox, K. Experience-dependent depression of vibrissae responses in adolescent rat barrel cortex. *Eur. J. Neurosci.* **10**, 2107–2116 (1998).
33. Fox, K. Anatomical pathways and molecular mechanisms for plasticity in the barrel cortex. *Neuroscience* **111**, 799–814 (2002).
34. Stern, E., Maravall, M. & Svoboda, K. Rapid development and plasticity of layer 2/3 maps in rat barrel cortex *in vivo*. *Neuron* **31**, 305–315 (2001).
35. Zhu, J. J. & Connors, B. W. Intrinsic firing patterns and whisker-evoked synaptic responses of neurons in the rat barrel cortex. *J. Neurophysiol.* **81**, 1171–1183 (1999).
36. Moore, C. I. & Nelson, S. B. Spatio-temporal subthreshold receptive fields in the vibrissa representation of rat primary somatosensory cortex. *J. Neurophysiol.* **80**, 2882–2892 (1998).
37. Huttenlocher, P. R., de Courten, C., Garey, L. J. & Van der Loos, H. Synaptogenesis in human visual cortex—evidence for synapse elimination during normal development. *Neurosci. Lett.* **33**, 247–252 (1982).
38. Rakic, P., Bourgeois, J. P., Eckenhoff, M. F., Zecevic, N. & Goldman-Rakic, P. S. Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science* **232**, 232–235 (1986).
39. Hubel, D. H. & Wiesel, T. N. Single-cell responses in striate cortex of kittens deprived of vision in one eye. *J. Neurophysiol.* **26**, 1003–1017 (1963).
40. Antonini, A. & Stryker, M. P. Rapid remodeling of axonal arbors in the visual cortex. *Science* **260**, 1819–1821 (1993).
41. Sandison, D. R., Piston, D. W. & Webb, W. W. *Three-Dimensional Confocal Microscopy: Volume Investigation of Biological Specimens* (eds Stevens, J. K., Mills, L. R. & Trogadis, J. E.) 29–47 (Academic, New York, 1994).
42. Harris, K. M. & Stevens, J. K. Dendritic spines of CA1 pyramidal cells in the rat hippocampus: serial electron microscopy with reference to their biophysical characteristics. *J. Neurosci.* **9**, 2982–2997 (1989).
43. Dunaevsky, A., Blazeski, R., Yuste, R. & Mason, C. Spine motility with synaptic contact. *Nature Neurosci.* **4**, 685–686 (2001).

Supplementary Information accompanies the paper on *Nature's* website (<http://www.nature.com/nature>).

Acknowledgements We thank M. Chklovskii for useful discussions, E. Ruthazer, W. Thompson, R. Weinberg and members of our laboratory for a critical reading of the manuscript, T. Pologruto and B. Sabatini for programming, and B. Burbach, P. O'Brien and A. Holtmaat for help with experiments. This work was supported by the Pew, Mathers, and Lehrman Foundations, HFSP and NIH (K.S.), and the Swiss National Science Foundation and HFSP (E.W.). B.C. is a predoctoral student at SUNY Stony Brook.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to K.S. (e-mail: svoboda@cshl.org).

Direct detection of variable tropospheric clouds near Titan's south pole

Michael E. Brown*, Antonin H. Bouchez* & Caitlin A. Griffith†

* Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA

† Lunar and Planetary Laboratory, University of Arizona, Tucson, Arizona 85721, USA

Atmospheric conditions on Saturn's largest satellite, Titan, allow the possibility that it could possess a methane condensation and precipitation cycle with many similarities to Earth's hydrological cycle. Detailed imaging studies^{1–4} of Titan have hitherto shown no direct evidence for tropospheric condensation clouds, although there has been indirect spectroscopic evidence for transient clouds^{5,6}. Here we report images and spectra of Titan that show clearly transient clouds, concentrated near the south pole, which is currently near the point of maximum solar heating. The discovery of these clouds demonstrates the existence of condensation and localized moist convection in Titan's atmosphere. Their location suggests that methane cloud formation is controlled seasonally by small variations in surface temperature, and that the clouds will move from the south to the north pole on a 15-year timescale.

Searching for clouds on Titan is best performed with high-resolution images and spectra at wavelengths between about 2.0 and 2.3 μm . At these wavelengths, absorption of photons by Titan's methane ranges from negligible (2.00 to 2.05 μm) to nearly complete (2.17 to 2.29 μm), while absorption by Titan's haze is low^{5,6}. The transparent range allows images at these wavelengths of the surface, while the complete range of methane absorption across the wavelength region allows spectra to probe all levels in Titan's atmosphere. High spatial resolution is necessary to recognize any small distinct cloud against the surface of Titan.

Using the adaptive optics (AO) system at the W. M. Keck observatory⁷, we have obtained images with both the spatial resolution and spectral coverage critical for searching for clouds. Adaptive optics achieves high spatial resolution by partially compensating for the smearing effect of turbulence in the Earth's atmosphere and delivering a near diffraction-limited image. Images from five nights of observations (Fig. 1) demonstrate the excellent spatial resolution achieved with the AO observations. Surface features which have been consistently imaged since 1995 are seen here at their highest resolution. The images from 10 and 11 December 2001 and 28 February 2002 all show almost the same face of Titan, and the reproducibility and rotation of surface features is apparent. More remarkable, however, are the transient changes visible near the south pole of Titan. These changes are more apparent in polar projections of the images (Fig. 2). The bright unresolved spot on the southern limb in the 10 and 11 December images has disappeared by February. The February image instead shows a much larger brightening extending about 1,400 km from 80° to 70° south latitude. Numerous small morphologically similar isolated bright spots occur throughout the images but cannot be verified to be transient because of a lack of duplicate coverage of most longitudes.

To determine the height in the atmosphere where the transient spots occur, we examined spectra from the 10 and 11 December bright southern spot and compared them to spectra obtained at a location 900 km directly east on the satellite, where the line-of-sight samples the same latitude at an identical zenith angle of 65° but no transient brightening occurs. A comparison between the two spectra

(Fig. 3) shows that they are essentially identical beyond 2.16 μm . These regions of the spectrum contain strong methane absorption lines which saturate in the stratosphere of the satellite. At shorter wavelengths the methane absorption is progressively weaker, and so photons penetrate progressively deeper into the atmosphere until at a wavelength of 2.12 μm photons reach the surface. The two spectra diverge at a wavelength of 2.155 μm , which means that there must be a reflective layer somewhere between the stratosphere and the surface. Detailed plane-parallel radiative transfer calculations using the models of Griffith *et al.*^{5,6} of the difference between the two spectra shows that the bright spot is best modelled as an unresolved highly reflective cloud layer with a filling factor of 25% at an altitude of 16 ± 5 km above Titan's surface, in the middle of Titan's troposphere. The errors in this height estimate are dominated by uncertainties in tropospheric methane abundance; deviation from the plane-parallel assumption at this zenith angle is less than 4%. No changes are apparent in the altitude, intensity or location of the cloud from 10 to 11 December. These images and spectra conclusively demonstrate the existence of transient clouds in the troposphere of Titan and point to the presence of a vigorous and currently active cycle of methane condensation and dissipation.

The December 2001 cloud has a brightness equivalent to about 0.3% of the total brightness of the disk of Titan at these wavelengths, and can be explained by a single (foreshortened) cloud of 200 km diameter or smaller clouds with the same total area. The 28 February 2002 cloud is significantly larger, reflecting a flux equivalent to about 1% of the total flux of Titan, and covering an apparent

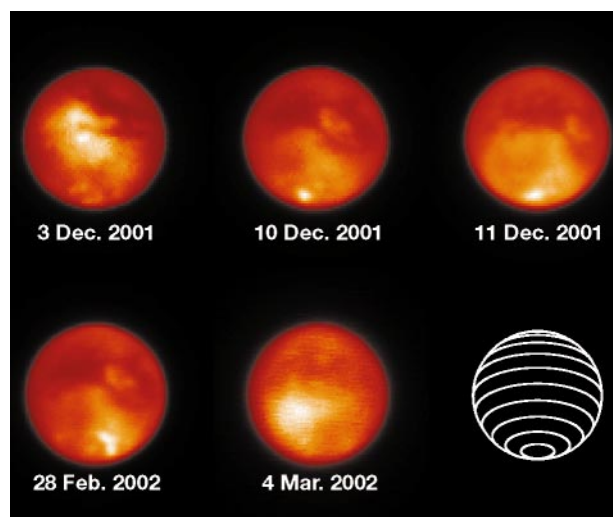


Figure 1 Images of Titan show the transient existence of cloud features near the south pole. Images and spectra from 3, 10 and 11 December 2001 were obtained using NIRSPEC, the facility near-infrared spectrograph¹⁸, while images (only) from 28 February and 4 March 2002 were obtained using NIRC2, the facility near-infrared AO imager. Images were obtained in the K' wavelength band, which extends from 1.96 to 2.99 μm , and have an angular resolution of 0.05 arcsec (a linear resolution of 330 km on 10 December 2001) on the satellite. The images shown are combinations of from 4 to 20 individual images shifted to a common centre, summed, and divided by an image of the individual pixel response function ('flat field'). The apparent elongation of the cloud feature on the 11 December 2001 image is a temporary artefact of the AO system. Owing to non-photometric observing conditions during some of the nights, no absolute flux calibration was obtained. The individual images are scaled to best see the polar clouds. The line figure shows every 15° of latitude projected for Titan's subsolar latitude of -25.6° at the time of the observations. Titan's subsolar longitudes at the times of observations were 69°, 228°, 249°, 235° and 325° for the 3, 10 and 11 December, and the 28 February and 4 March observations, respectively.

area of $4.4 \times 10^5 \text{ km}^2$, which implies a filling factor of only about 5%.

These transient clouds will cause rapid changes to Titan's full-disk spectrum similar to the approximately 0.5% variations previously observed shortward of $2.170 \mu\text{m}$ and interpreted as evidence for such clouds⁶. The difference in wavelength of spectral divergence between $2.170 \mu\text{m}$ in the full-disk spectra and $2.155 \mu\text{m}$ in the current spectra suggests a change in cloud height or latitude since the time of the previous observations.

No obvious connection exists between these transient tropospheric clouds and the southern tropopause level scattering layer^{8–11} which occurs at an altitude of around 40 km and is visible in the spectra of Fig. 3. This layer has been observed since at least 1994 and has been described as a cloud, haze layer or fog, but its true physical nature, composition and cause remain unknown. Although this scattering layer is routinely observed, the lack of a consistent name has caused confusion as to its stability and even existence. We use the term 'tropopause cirrus' as a morphological name for this distinct layer which describes its approximate altitude, small optical depth and large areal coverage. At the time of observations, this cirrus layer covered the entire southern hemisphere at latitudes further south than 35° (ref. 11). The southern tropopause cirrus layer is not currently known to vary, although it is likely to change on seasonal timescales with changes in stratospheric haze chemistry and dynamics⁹.

The most striking property of these transient cloud events is their unexpected concentration near the south pole of Titan. Although

heating at southern summer solstice might be expected to drive polar convection, studies of tropospheric conditions on Titan have suggested an absence of seasonal variation^{12,13} and predicted that methane clouds, if present, should concentrate at the equator year-round¹⁴. These hypothesized seasonal invariances come from consideration of the long radiative timescale of Titan's lower troposphere which does not allow tropospheric temperatures to appreciably change on seasonal timescales¹².

These predictions of invariance do not consider the effects of small seasonally varying surface temperatures, however, and instead examine only conditions measured during the northern spring equinox at the time of the Voyager fly-by. Even a very small seasonally and latitudinally varying surface temperature, which is

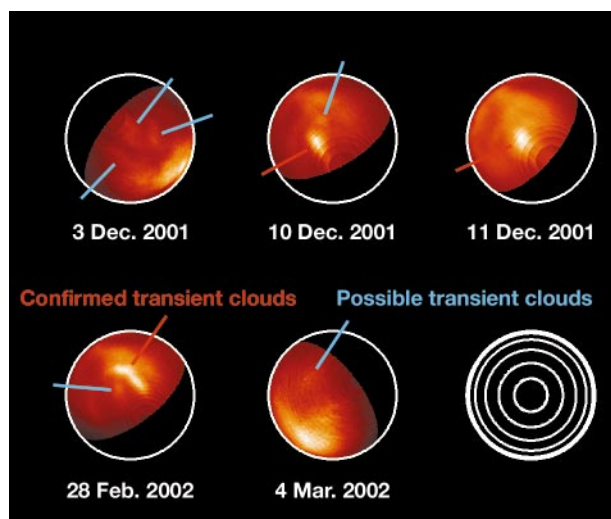


Figure 2 The transient polar clouds are best seen in polar projections of the images of Titan. A longitude of zero (facing Saturn) is straight down in these projections and longitude increases anticlockwise. The line figure shows every 15° of latitude. Confirmed transients (those for which we have several images confirming change) are labelled by red lines, while morphologically similar features whose transient nature is unconfirmed are labelled by blue lines. On 10 and 11 December, the single brightest spot appears at long. $150^\circ \pm 30^\circ$, lat. $77^\circ \pm 3^\circ$ south (the large errors in locating spots near the limb precludes any detection of cloud motion between the two nights). The 28 February images show small brightenings at long. $200^\circ \pm 20^\circ$, lat. $87^\circ \pm 5^\circ$ south and long. $174^\circ \pm 10^\circ$, lat. $65^\circ \pm 5^\circ$ south, but none at the precise location of that seen in December. Also visible is a much larger brightening between latitudes 70° and 80° south; nothing similar is seen in the December images. The projections were created by using the circular symmetry of the limb of Titan to define the centre of the disk and then projecting the intensity onto a spherical grid. Errors in the position of the centre of the disk of ± 0.02 arcsec give errors in the projected longitude and latitude of features near the south pole of 20° and 5° respectively. No intensity corrections are made for variations in the viewing angle.

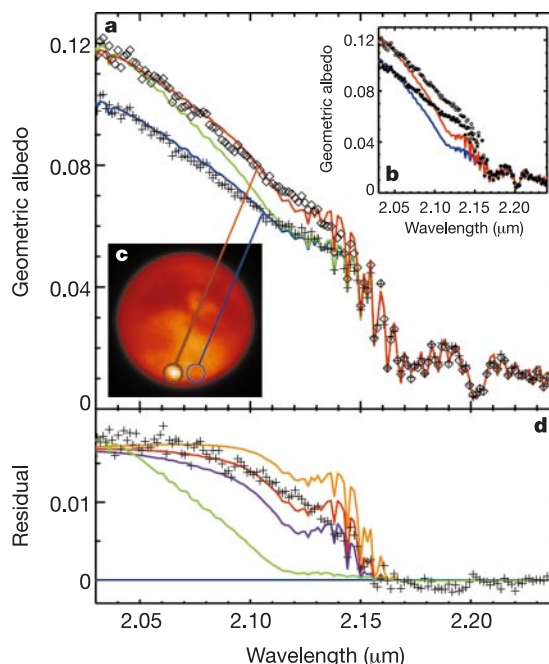


Figure 3 Spatially resolved spectra of the disk of Titan showing the height in the atmosphere of the transient features. **a**, A comparison of a spectrum of the 10 December 2002 transient brightening and of a location 900 km east on the disk (locations shown in **c**), which samples the surface and atmosphere at an identical zenith angle of 65° , shows that the transient brightening is caused by tropospheric clouds. **a** compares spectra from the regions labelled on the 10 December image with best-fit radiative transfer models. The blue line shows the best-fit model to the non-cloud region, while the red line shows the effect of adding a bright scattering layer at 18 km. The green line demonstrates the poor fit obtained if an attempt is made to fit the transient brightening as a surface albedo effect. Neither spectrum can be modelled without the presence of a tropopause cirrus layer of optical depth 0.10 ± 0.02 at $2 \mu\text{m}$ at 30–40 km (refs 8–11). **b** shows the best-fit models without this tropopause cirrus. **d** shows the difference between the two spectra and between models with clouds at 14 (purple), 18 (red), and 24 km (orange). The green line again shows the poor fit if the spectral difference is modelled purely with a change in surface albedo. The mismatches between the models and data are due to inaccuracies in the methane line strengths; these inaccuracies contribute the greatest uncertainty to the derived cloud height. The spectra were obtained simultaneously using a 0.054 arcsecond wide spectral slit centred on the brightening and placed east–west across the disk. Both spectra have been divided by the spectrum of the G4V star HD32923 taken at an identical airmass of 1.02 to correct for effects of instrumental sensitivity, telluric absorption, and solar spectrum. The radiative transfer model uses doubling and adding techniques to approximate the equation of radiative transfer, along with line-by-line and Mie scattering calculations to incorporate gas absorption and scattering by particulates. The nominal value of methane humidity is that of ref. 6; changes of factors of two in methane or allowing methane supersaturation do not significantly change the results.

a necessary consequence of Titan's widely varying seasonal insolation, will affect the magnitude of the surface sensible heat flux and change temperatures in the lower troposphere. If the lower tropospheric lapse rate is close to the boundary between stability and instability, as has been measured on Titan^{13,15,16}, a small additional heat flux can drive the creation of a thermally convective layer, the height of which will depend on the magnitude of the additional heat input and therefore on the surface temperature. Assuming the conditions least favourable to convection that have been inferred from Voyager measurements¹³, even a surface temperature rise of only 1 K is sufficient to cause a convective layer 7 km in height. If this convective layer reaches the point at which methane saturates, the height of which is highly dependent on the methane humidity but is estimated to be around 4 km in typical models^{6,12,15,16}, methane condensation will render the air buoyant and drive clouds to the levels at about 15 km observed¹⁷. In regions with lower surface temperatures, the convective layer will be smaller or even non-existent, and condensation will not occur.

At the time of our observations, Titan was approaching southern summer solstice, and, owing to Titan's obliquity of 27°, the polar regions were in continuous sunlight and receiving more daily averaged insolation than any other spot on the satellite (and 50% more daily averaged insolation than the equator at equinox). We suggest that this insolation leads to a maximum surface temperature in these polar regions which drives a convective layer large enough to cause methane condensation and the ensuing moist convection. This hypothesis predicts that the location of these convective clouds will follow the location of maximum insolation (with some lag owing to the thermal inertia of the surface). □

Received 31 July; accepted 19 November 2002; doi:10.1038/nature01302.

- Smith, P. H. *et al.* Titan's surface, revealed by HST imaging. *Icarus* **119**, 336–349 (1996).
- Combes, M. *et al.* Spatially resolved images of Titan by means of adaptive optics. *Icarus* **129**, 482–497 (1997).
- Gibbard, S. G. *et al.* Titan: High-resolution speckle images from the Keck telescope. *Icarus* **139**, 189–201 (1999).
- Cousteins, A. *et al.* Images of Titan at 1.3 and 1.6 μm with adaptive optics at the CFHT. *Icarus* **154**, 501–515 (2001).
- Griffith, C. A., Owen, T., Miller, G. A. & Geballe, T. Transient clouds in Titan's lower atmosphere. *Nature* **395**, 575–578 (1998).
- Griffith, C. A., Hall, J. L. & Geballe, T. R. Detection of daily clouds on Titan. *Science* **290**, 509–512 (2000).
- Wizinowich, P. *et al.* Performance of the W.M. Keck Observatory natural guide star adaptive optics facility: the first year at the telescope. *Proc. SPIE* **4007**, 2–13 (2000).
- Bouchez, A. H., Brown, M. E., Griffith, C. A. & Dekany, R. G. Spatially resolved spectroscopy of Titan's surface and atmosphere. *Bull. Am. Astron. Soc.* **32**, 1702 (2000).
- Lorenz, R. D., Young, E. F. & Lemmon, M. T. Titan's smile and collar: HST observations of seasonal change 1994–2000. *Geophys. Res. Lett.* **28**, 4453–4456 (2001).
- Griffith, C. A., Hall, J. L., Young, E., Cook, J. & Rannou, P. Imaging temporal changes on Titan. *Icarus* (in the press).
- Bouchez, A. H., Brown, M. E., Neyman, C. R., Troy, M. & Dekany, R. G. Ground-based visible adaptive optics observations of Titan's atmosphere and surface. *Icarus* (submitted).
- Hunten, D. M. *et al.* *Saturn* (eds Gehrels, T. & Matthews, M. S.) 671–759 (Univ. Arizona Press, Tucson, 1984).
- McKay, C. P., Martin, S. C., Griffith, C. A. & Keller, R. M. Temperature lapse rate and methane in Titan's troposphere. *Icarus* **129**, 498–505 (1997).
- Tokano, T., Neubauer, F. M., Laube, M. & McKay, C. P. Three-dimensional modeling of the tropospheric methane cycle on Titan. *Icarus* **153**, 130–147 (2001).
- Lindal, G. F. *et al.* The atmosphere of Titan: An analysis of the Voyager 1 radio-occultation measurements. *Icarus* **53**, 348–363 (1983).
- Lellouch, E. *et al.* Titan's atmosphere and hypothesized ocean: A reanalysis of Voyager 1 radio-occultation and IRIS 7.7- μm data. *Icarus* **79**, 328–349 (1989).
- Awal, M. & Lunine, J. I. Moist convective clouds in Titan's atmosphere. *Geophys. Res. Lett.* **21**, 2491–2494 (1994).
- McLean, I. S. *et al.* The design and development of NIRSPEC: a near-infrared echelle spectrograph for the Keck II telescope. *Proc. SPIE* **3354**, 566–578 (1998).

Acknowledgements We thank E.J. Moyer and M.I. Richardson for conversations, D. Le Mignaut, R. Campbell, M. Konacki and J. Eisner for acquiring the NIRC2 data, and S. Hörst for many nights of monitoring Titan in the cold. This work was supported by a grant from the NSF Planetary Astronomy programme.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.E.B. (e-mail: mbrown@caltech.edu).

Charge-ordered ferromagnetic phase in $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$

James C. Loudon, Neil D. Mathur & Paul A. Midgley

Department of Materials Science and Metallurgy, University of Cambridge, Pembroke Street, Cambridge CB2 3QZ, UK

Mixed-valent manganites are noted for their unusual magnetic, electronic and structural phase transitions. For example, the $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ phase diagram¹ shows that below transition temperatures in the range 100–260 K, compounds with $0.2 < x < 0.5$ are ferromagnetic and metallic, whereas those with $0.5 < x < 0.9$ are antiferromagnetic and charge ordered. In a narrow region around $x = 0.5$, these totally dissimilar ground states are thought to coexist^{2,3}. It has been shown⁴ that charge order and charge disorder can coexist in the related compound, $\text{La}_{0.25}\text{Pr}_{0.375}\text{Ca}_{0.375}\text{MnO}_3$. Here we present electron microscopy data for $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ that shed light on the distribution of these coexisting phases, and uncover an additional, unexpected phase. Using electron holography and Fresnel imaging, we find micrometre-sized ferromagnetic regions spanning several grains coexisting with similar-sized regions with no local magnetization. Holography shows that the ferromagnetic regions have a local magnetization of 3.4 ± 0.2 Bohr magnetons per Mn atom (the spin-aligned value is $3.5 \mu_B$ per Mn). We use electron diffraction and dark-field imaging to show that charge order exists in regions with no net magnetization and, surprisingly, can also occur in ferromagnetic regions.

Figure 1 shows the temperature dependence of the magnetization for three different compositions of polycrystalline $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ with a grain size of $5 \mu\text{m}$ (produced by Praxair, Woodinville, Washington, USA). These measurements were made in a field of 1 T with a vibrating sample magnetometer. The $x = 0.3$ compound (Fig. 1a) shows a standard paramagnetic-to-ferromagnetic phase transition at 256 K, and approaches the theoretical saturation moment of $3.7 \mu_B$ per Mn (derived from the filling of the Mn d levels) at low temperatures (down to 10 K) as expected. The magnetization of the $x = 0.66$ compound (Fig. 1c) shows a paramagnetic-to-antiferromagnetic phase transition at 250 K. The $x = 0.5$ sample (from which the electron microscopy data is taken) appears to show both a paramagnetic-to-ferromagnetic transition at 233 K, and then an antiferromagnetic transition at 120 K measured on the cooling curve and 175 K on the warming curve (Fig. 1b). The large thermal hysteresis has been reported by other authors^{1,5}, and may be an indication that charge ordering is a 'nucleation and growth' process. Below the Néel temperature, the magnetization is still higher than at room temperature, and there are two possible causes of this effect: spin canting^{6,7}, or an inhomogeneous mixture of ferromagnetic and antiferromagnetic regions^{2,4} within the same sample. We show here that the latter is the correct explanation.

According to the conventional models^{8–10} for the low-temperature phase transitions in manganites, the antiferromagnetic phase is associated with a spontaneous ordering of Mn^{3+} and Mn^{4+} ions. If the room-temperature cell is indexed¹¹ as orthorhombic $Pnma$ ($a = 5.42 \text{ \AA}$, $b = 7.65 \text{ \AA}$, $c = 5.44 \text{ \AA}$), the lattice distortion caused by this charge ordering gives rise to extra reflections at positions $(h + q, k, l)$ at low temperatures¹² in a diffraction pattern where q is the wavevector of the modulation. The wavevector has been seen to vary with both composition and temperature¹³, but here we consider only the commensurate phase $q = 1/2$. If an objective aperture is placed over one of these reflections, parts of the sample that charge order appear bright in an image.

The $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ sample was prepared for transmission electron microscopy by conventional mechanical polishing and argon-ion thinning at liquid-nitrogen temperatures. Large-area ($10 \times 10 \mu\text{m}$) energy-dispersive X-ray analysis using a scanning electron microscope showed that the La/Ca ratio was constant to within the error of the measurements ($\delta x = \pm 2\%$). Transmission electron microscopy was undertaken using a 300-kV Philips CM300 microscope equipped with a field-emission gun to provide a coherent source of electrons, a biprism used for holography and a 'Lorentz' lens (a strong objective mini lens), which allows the magnetic structure of a specimen to be observed in near-zero (less than 0.02 T) magnetic field. A liquid-nitrogen specimen stage was used to cool the samples to 90 K.

We used two methods to image the magnetic structure of the material: Fresnel imaging and electron holography. In the Fresnel method¹⁴, the sample is imaged out of focus with a coherent beam of electrons. Lorentz forces in the sample cause magnetic domain walls to appear¹⁵ as bright interference fringes where the electrons converge, and as dark regions where the electrons diverge. In electron holography¹⁶, an electron biprism (a positively charged wire) is inserted into the column of the microscope, and used to interfere a

reference wave passing through vacuum with one passing through the sample. A digital reconstruction allows both the amplitude and phase of the exit wave-function to be determined directly.

Taking coordinates x and y normal to the beam direction z , the phase change on passing through the specimen is given by¹⁶

$$\phi(x, y) = C_E \int V_0(x, y, z) dz - \frac{2\pi e}{h} \iint \mathbf{B}_\perp(x, y, z) \cdot d\mathbf{S} \quad (1)$$

where $C_E = \frac{2\pi e}{\lambda} \left(\frac{eV + mc^2}{eV(eV + 2mc^2)} \right) = 6.523 \times 10^6 \text{ m}^{-1} \text{ V}^{-1}$, e is the electron charge, m is the electron rest mass, V is the acceleration voltage (300 kV), c is the speed of light, λ is the wavelength of an electron (0.0197 Å at 300 kV), $V_0(x, y, z)$ is the mean inner potential, $\mathbf{B}_\perp(x, y, z)$ is the component of magnetic flux density normal to the electron beam, and $d\mathbf{S}$ is an element of vector area normal to the beam direction. This formula holds true provided that the specimen is not diffracting strongly. This is ensured by tilting away from any major zone axes.

By comparing holograms taken at room temperature (above the Curie temperature, T_C) with those taken at 90 K (below T_C), we can rearrange equation (1) to give a measure of the absolute value of the magnetization in a way that does not depend on the unknown thickness of the sample:

$$\begin{aligned} \mu_0 \mathbf{M}_\perp(x, y) &= \mathbf{B}_\perp(x, y) \\ &= -\frac{h}{2\pi e} C_E V_0 \frac{1}{\phi_{T > T_C}(x, y)} \\ &\quad \times \left(\frac{\partial/\partial y}{-\partial/\partial x} \right) \{ \phi_{T < T_C}(x, y) - \phi_{T > T_C}(x, y) \} \end{aligned} \quad (2)$$

This formula assumes that: (1) V_0 is constant throughout the sample and has the value of 23.73 V calculated from electron scattering factors¹⁷, (2) $\mathbf{B}_\perp(x, y)$ is constant throughout the thickness of the sample, and (3) the stray field is negligible in comparison with the flux density within the sample.

In Fig. 2 we show ferromagnetism and charge order coexisting. The field of view of a hologram is limited by the size of the interference region created by the biprism, and so we present Fig. 2a as a montage of regions investigated by holography (colour) overlaid on a Fresnel image. This image was taken using the Lorentz lens to focus the electrons. Typically, this lens places the specimen in a vertical magnetic field of less than 0.02 T. (Another Fresnel image showing domain walls in grain 2 is presented in Supplementary Fig. 1). The Fresnel image shows long straight black and white lines in the lower left-hand corner. These are magnetic domain walls, which match the coloured domains derived from holography. (The black curves at the top right are bend contours.) Three grains are present in Fig. 2a, labelled 1, 2 and 3. In grain 1, the orthorhombic $\mathbf{b}$ axis lies approximately in the plane and 180° domains dominate, whereas in grain 2, the $\mathbf{b}$ axis lies approximately normal to the film and there is a mixture of 90° and 180° domains. This change in magnetic behaviour probably arises because of the change in the orientation of the unit cell and the shape anisotropy caused by the thin specimen¹⁸.

The magnetization measured over $100 \times 100 \text{ nm}$ squares at several positions in several domains in grains 1 and 2 was $3.4 \mu_B$ per Mn on average, with a standard deviation of $0.5 \mu_B$ per Mn. If it is assumed that this variability comes only from random errors, we can give the average local magnetization as $3.4 \pm 0.2 \mu_B$ per Mn. The holograms showed that the stray field in the vacuum near the specimen edge was approximately 5% of the flux density in the sample, so the stray field may be safely neglected. The mean inner potential V_0 was derived from electron scattering factors, assuming that all the atoms are neutral and no bonding takes place. This assumption is likely to overestimate the magnetization, possibly by

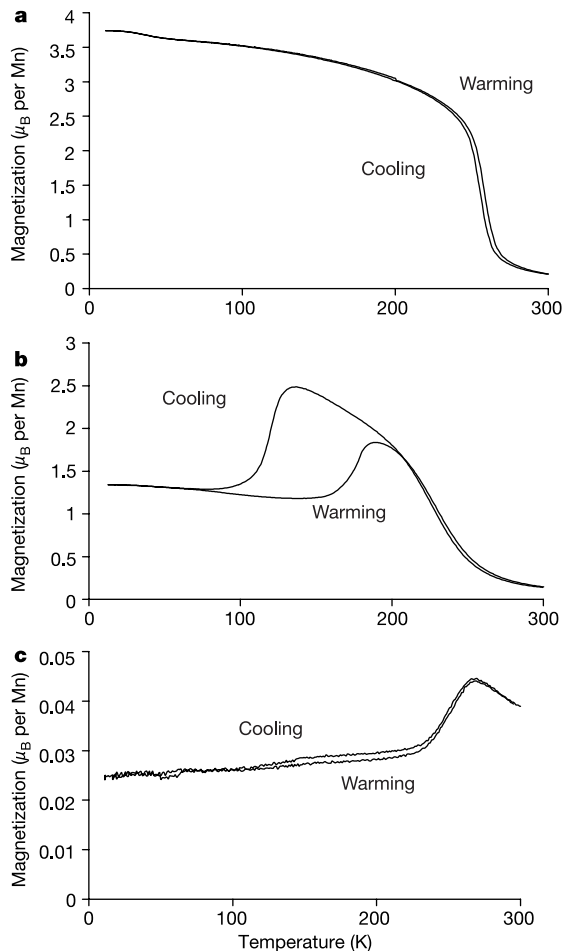


Figure 1 Curves of magnetization against temperature for various compositions of polycrystalline $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ under an applied field of 1 T. **a**, $x = 0.3$ shows a conventional paramagnetic-to-ferromagnetic phase transition (with a Curie temperature of 256 K measured at the inflexion point), and approaches the saturation moment of $3.7 \mu_B$ per Mn. The small hysteresis is almost certainly due to thermal inertia in the cryostat. **b**, $x = 0.5$ appears to show two transitions as the temperature is lowered: first, a paramagnetic-to-ferromagnetic transition at 233 K, and second, an antiferromagnetic transition at 120 K measured on the cooling curve and 175 K on the warming curve. **c**, $x = 0.66$ shows a paramagnetic-to-antiferromagnetic transition at 250 K.

as much as 10%. In comparing this magnetization with the saturation moment, we assume that the magnetization lies in the plane of the specimen as would be expected by considering the shape anisotropy. The magnetization of $3.4 \pm 0.2 \mu_B$ per Mn measured here agrees well with the value of $3.5 \mu_B$ per Mn that would be expected if the sample were fully ferromagnetic, and is significantly larger than the bulk value of $1.25 \mu_B$ per Mn measured by magnetometry.

Holography showed that grain 3 had no net magnetization, as did several other grains in other regions of the specimen. Figure 2e shows a diffraction pattern taken from the circled region in grain 3 showing charge ordering. This is the expected charge-ordered, non-ferromagnetic phase. We conclude that at 90 K, the sample is composed of an inhomogeneous mixture of ferromagnetic regions and regions with no net magnetization. Given that the bulk magnetization at this temperature is $1.25 \mu_B$ per Mn (compared with the theoretical value of $3.5 \mu_B$ per Mn), around 36% of the total sample is ferromagnetic.

Figure 2b shows a dark-field image from grain 2 taken using the $(1/2, 0, 2)$ (charge-ordered) reflection, and Fig. 2c shows the associated diffraction pattern taken from the region circled in grain 2, showing charge-ordered reflections at positions $(h + 1/2, k, l)$. These images were taken with the Lorentz lens turned off and the electrons focused with the main objective lens, which places the sample in a vertical magnetic field of around 2 T. This will tend to cause the magnetization to rotate towards the optic axis, and

possibly cause the charge-ordered regions to shrink. Despite this, both the diffraction pattern and the dark-field image show charge order present in the region of the sample where holography revealed ferromagnetic domains. The simplest explanation for the dark-field contrast seen in grain 2 (Fig. 2b) is that the bright regions are charge ordered and ferromagnetic, and the dark regions are ferromagnetic and charge disordered.

We now consider the possibility that the ferromagnetic charge-ordered phase is in fact made of charge-ordered antiferromagnetic regions coexisting with ferromagnetic metallic regions on a very small scale. The spatial resolution of the dark-field image is determined primarily by the size of the objective aperture and the specimen drift rate, and is about 2 nm in this case. The spatial resolution of the holograms is limited by the size of the window used in the Fourier filtering process as part of the reconstruction process and is about 20 nm; the system can therefore resolve the magnetization and its distribution in any region several times larger than this. The large bright patch in the dark-field image (Fig. 2b), which overlaps the magnetic region in Fig. 2a, is about 100 nm in size. Thus if a small-scale segregation takes place, the coexisting phases must be less than 2 nm (about four unit cells parallel to the *a* axis) in size; this must be regarded as mesoscopic texture rather than two thermodynamically stable phases as argued in ref. 3. Furthermore, if this were the case, we would expect the magnetization in the overlapping region to be approximately halved. In fact, the magnetization in this region is $3.5 \pm 0.2 \mu_B$ per Mn, in excellent agreement with the spin-

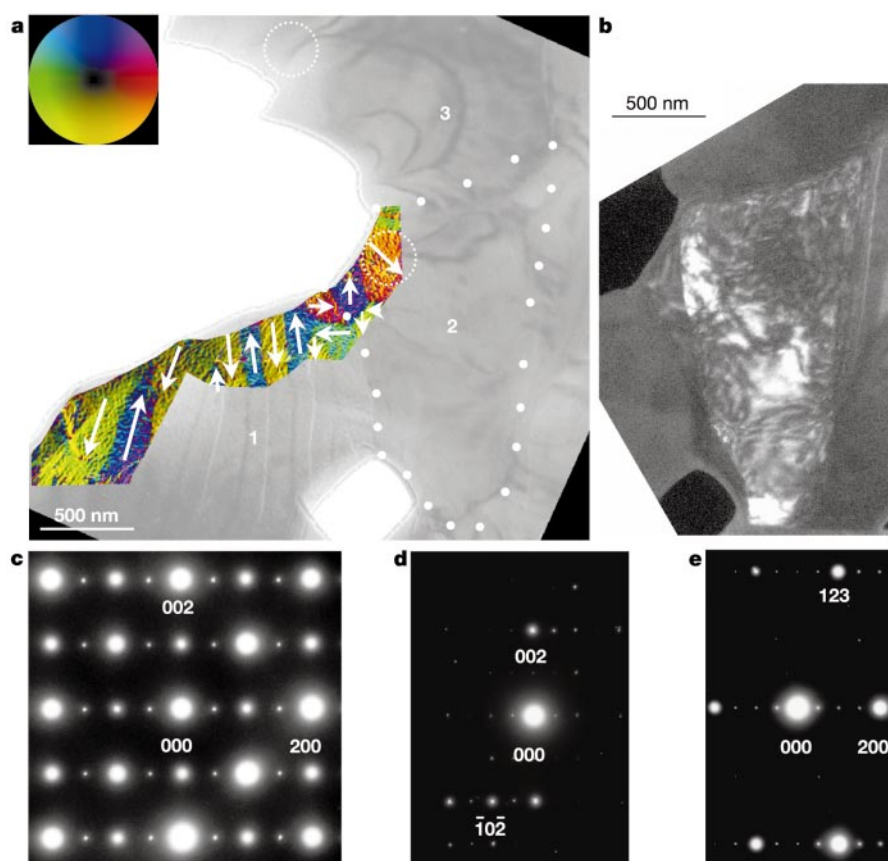


Figure 2 Coexistence of ferromagnetism and charge order. Three grains (labelled 1, 2 and 3) are present in the image. The grain boundary is indicated by the white dots. **a**, A colour image of the magnetic structure derived from available holography data overlaid on a Fresnel image. The colour wheel indicates the direction of the local magnetization (marked with arrows). The apparent texture within each domain is likely to be an artefact of the Fourier filtering process used to reconstruct the phase of the exit wavefunction. **b**, A dark-field image on the same scale from grain 2, taken using the $(1/2, 0, 2)$ charge-ordered reflection. **c**, A diffraction pattern taken from the circled

region in 2 looking down the $[010]$ axis with charge-ordered reflections at $(h + 1/2, k, l)$. **d**, A diffraction pattern from the same region as **c** tilted away from the $[010]$ zone axis by a few degrees. The charge-ordered reflections persist, but reflections of type $(00l)$ (for *l* odd) are now absent as they are kinematically forbidden by the *Pnma* space group. **e**, A diffraction pattern from the circled region in grain 3 looking down the $[032]$ zone axis. Charge-ordered reflections can be seen at $(h + 1/2, k, l)$. Holography showed that this region had no net magnetization. The very weak reflections between the main rows are from a twin variant of this axis, the $[214]$ zone axis.

aligned value of $3.5 \mu_B$ per Mn. It is also possible that this region may be a ferromagnetic phase on top of a separate non-ferromagnetic, charge-ordered phase. If so, the charge-ordered phase would have to be less than 2 nm (about two unit cells parallel to the **b** axis) thick to produce a magnetization within the error limits quoted. Such a thin region is unlikely, because the region of interest is very bright in the dark-field image (Fig. 2b), indicating that charge order is well established here. We have also taken a number of Fresnel images and diffraction patterns from other areas of the sample, and have seen magnetic domain walls and charge ordering in the same region (Supplementary Fig. II), which is further evidence for this unexpected phase.

In tilting the sample, we found that charge-ordered reflections perpendicular to a systematic row were far stronger for those parallel to the systematic row. Figure 2d shows a systematic row from the same region as Fig. 2c with strong charge-ordered reflections perpendicular to the row. This is strong evidence that charge order is predominantly a transverse modulation of the atoms, as discussed in ref. 11 based on evidence from neutron powder diffraction.

Using electron microscopy, we have shown that at 90 K, $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ forms an inhomogeneous mixture of ferromagnetic and zero-moment (presumably antiferromagnetic¹) regions. Each region extends for several micrometres, and can span several crystallographic grains. Our results suggest that charge order occurs not only in regions with no net magnetization, but can also occur in ferromagnetic regions. Very recently, evidence has been published¹⁹ that is consistent with the possibility of a similar coexistence in $\text{La}_{0.25}\text{Pr}_{0.375}\text{Ca}_{0.375}\text{MnO}_3$. This charge-ordered, ferromagnetic phase is not predicted by conventional models^{8–10}, but unexpected phases such as this can arise from a simple hamiltonian²⁰.

We suggest that the bandgap that opens up below the charge-ordering transition temperature can be small enough that the valence electrons still have enough mobility to promote ferromagnetism via double exchange, and yet large enough to produce a charge density wave. This seems reasonable, as ferromagnetic coupling in the manganites should only require nearest-neighbour hopping. Furthermore, some degree of charge ordering may even be possible deep within the ferromagnetic regime, as diffuse charge-ordered reflections have been observed²¹ in $\text{La}_{2/3}\text{Ca}_{1/3}\text{MnO}_3$. We further suggest that the bandgap is locally modified by strain within each grain, and that this results in the rich variety of phases that can coexist in the manganites. This may also explain the strong dependence of $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$ properties on grain size⁵. Therefore manganite phase diagrams based on composition alone paint an inadequate picture of manganite physics, and miss the richness and complexity that we have demonstrated here. □

Received 18 June; accepted 14 November 2002; doi:10.1038/nature01299.

- Cheong, S.-W. & Hwang, H. Y. *Colossal Magnetoresistive Oxides* Monographs in Condensed Matter Science, Ch. 7 (ed. Tokura, Y.) (Gordon & Breach, Reading, UK, 2000).
- Moreno, A., Yunoki, S. & Dagotto, E. Phase separation scenario for manganese oxides and related materials. *Science* **283**, 2034–2040 (1999).
- Mathur, N. D. & Littlewood, P. B. The self-organised phases of manganites. *Solid State Commun.* **119**, 271–280 (2001).
- Uehara, M., Mori, S., Chen, C. H. & Cheong, S.-W. Percolative phase separation underlies colossal magnetoresistance in mixed valent manganites. *Nature* **399**, 560–563 (1999).
- Levy, P. et al. Controlled phase separation in $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$. *Phys. Rev. B* **62**, 6437–6441 (2000).
- Radelli, P. G., Cox, D. E., Capogna, L., Cheong, S.-W. & Marezio, M. Wigner-crystal and bi-stripe models for the magnetic and crystallographic superstructures of $\text{La}_{0.333}\text{Ca}_{0.667}\text{MnO}_3$. *Phys. Rev. B* **59**, 14440–14450 (1999).
- Roy, M., Mitchell, J. F., Ramirez, A. P. & Schiffer, P. Doping-induced transition from double exchange to charge order in $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ near $x = 0.5$. *Phys. Rev. B* **58**, 5185–5188 (1998).
- Goodenough, J. B. Theory of the role of covalence in the perovskite type manganites $[\text{La}(\text{M}(\text{II}))\text{MnO}_3]$. *Phys. Rev.* **100**, 564–573 (1955).
- De Gennes, P.-G. Effects of double exchange in magnetic crystals. *Phys. Rev.* **118**, 141–154 (1960).
- Zener, C. Interaction between the d shells in the transition metals. II. Ferromagnetic compounds of manganese with perovskite structure. *Phys. Rev.* **82**, 403–405 (1951).
- Radelli, P. G., Cox, D., Marezio, M. & Cheong, S.-W. Charge, orbital and magnetic ordering in $\text{La}_{0.5}\text{Ca}_{0.5}\text{MnO}_3$. *Phys. Rev. B* **55**, 3015–3023 (1997).
- Mori, S., Chen, C. H. & Cheong, S.-W. Pairing of charge-ordered stripes in $(\text{La}, \text{Ca})\text{MnO}_3$. *Nature* **392**, 473–476 (1998).

- Chen, C. H., Cheong, S.-W. & Hwang, H. Y. Charge-ordered stripes in $\text{La}_{1-x}\text{Ca}_x\text{MnO}_3$ with $x > 0.5$. *J. Appl. Phys.* **81**, 4326–4330 (1997).
- Hirsh, P., Howie, A., Nicholson, R., Pashley, D. W. & Whelan, M. J. *Electron Microscopy of Thin Crystals*, 2nd edn (Krieger, Malabar, Florida, 1977).
- Lloyd, S. J., Mathur, N. D., Loudon, J. C. & Midgley, P. A. Magnetic domain wall width in $\text{La}_{0.7}\text{Ca}_{0.3}\text{MnO}_3$ thin films using Fresnel imaging. *Phys. Rev. B* **64**, 172407 (2001).
- Dunin-Borkowski, R. E., McCartney, M. R., Smith, D. J. & Parkin, S. S. P. Towards quantitative electron holography of magnetic thin films using in-situ magnetization reversal. *Ultramicroscopy* **74**, 61–73 (1998).
- Rez, D., Rez, P. & Grant, I. Dirac-Fock calculations of x-ray scattering factors and contributions to the mean inner potential for electron scattering. *Acta Crystallogr. A* **50**, 481–497 (1994).
- Suzuki, Y. et al. Magnetic anisotropy of doped manganite thin films and crystals. *J. Appl. Phys.* **83**, 7064–7066 (1998).
- Mori, S., Asaka, T. & Matsui, Y. Observation of magnetic domain structure in phase-separated manganites by Lorentz electron microscopy. *J. Electron Microsc.* **51**, 225–229 (2002).
- Yunoki, S., Hotta, T. & Dagotto, E. Ferromagnetic A-type, and charge-ordered CE-type states in doped manganites using Jahn-Teller phonons. *Phys. Rev. Lett.* **84**, 3714–3717 (2000).
- Zuo, J. M. & Tao, J. Nanometer-sized regions of charge ordering and charge melting in $\text{La}_{2/3}\text{Ca}_{1/3}\text{MnO}_3$ revealed by electron microdiffraction. *Phys. Rev. B* **63**, 060407 (2001).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank R. Dunin-Borkowski for help with electron holography, D. Vowles for help in using the scanning electron microscope, A. Williams and J. P. Attfield for discussions, and A. J. Millis and P. B. Littlewood for information on the nature of charge order. This work was funded by the UK EPSRC and by the Royal Society.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.C.L. (e-mail: james.loudon@physics.org).

Electroluminescence from single monolayers of nanocrystals in molecular organic devices

Seth Coe^{*†}, Wing-Keung Woo^{†‡}, Mouni Bawendi[‡] & Vladimir Bulović^{*}

^{*} Laboratory of Organic Optoelectronics, Department of Electrical Engineering and Computer Science,

[‡] Center for Materials Science and Engineering, Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[†] These authors contributed equally to this work

The integration of organic and inorganic materials at the nanometre scale into hybrid optoelectronic structures enables active devices^{1–3} that combine the diversity of organic materials with the high-performance electronic and optical properties of inorganic nanocrystals⁴. The optimization of such hybrid devices ultimately depends upon the precise positioning of the functionally distinct materials. Previous studies^{5,6} have already emphasized that this is a challenge, owing to the lack of well-developed nanometre-scale fabrication techniques. Here we demonstrate a hybrid light-emitting diode (LED) that combines the ease of processability of organic materials with the narrow-band, efficient luminescence of colloidal quantum dots⁷ (QDs). To isolate the luminescence processes from charge conduction, we fabricate a quantum-dot LED (QD-LED) that contains only a single monolayer of QDs, sandwiched between two organic thin films. This is achieved by a method that uses material phase segregation between the QD aliphatic capping groups and the aromatic organic materials. In our devices, where QDs function exclusively as lumophores, we observe a 25-fold improvement in luminescence efficiency (1.6 cd A^{-1} at $2,000 \text{ cd m}^{-2}$) over the best previous QD-LED results⁵. The reproducibility and precision of our

phase-segregation approach suggests that this technique could be widely applicable to the fabrication of other hybrid organic/inorganic devices.

The development of solution-based synthetic routes for creating CdSe(ZnS) nanocrystal QDs generated a material set that was continuously tunable, narrow-band, and an efficient emitter of saturated colour across the visible spectrum⁷. Solution photoluminescence quantum efficiencies of QDs can exceed 50% (ref. 8), and by controlling the QD diameter during synthesis, the peak emission wavelength can be tuned continuously from 470 nm to 630 nm. Colloidal QD solution samples can be routinely prepared with size distributions of less than 6% standard deviation, corresponding to a photoluminescence full-width at half-maximum (FWHM) of less than 30 nm. These desirable emissive properties, unmatched by any class of organic chromophores, motivated the creation of hybrid organic/inorganic LEDs containing QDs^{1,5}. To date the demonstrated efficiencies of QD-LEDs are far below those of state of the art all-organic LED technology^{9–11}. However, investigation into the fundamental limits of QD-LED performance reveals their potential to become a versatile technological platform for the creation of planar light emitters. All of this motivates our own work in redesigning hybrid device structures on the nanoscale, in the hope of creating QD-LEDs that have properties unattainable using only organic materials, and efficiencies that make them technologically viable.

To isolate the role of QDs in the LED luminescence process from their participation in charge conduction, we fabricate hybrid QD-LEDs containing only a single monolayer of QDs sandwiched between organic thin films. The organic layers transport charge carriers to the vicinity of the QD monolayer from which the luminescence originates. This is in contrast to previous studies that utilized QD multilayer films, of the order of 10–20 layers thick, which had the dual function of both transporting electrons and serving as the emissive layer^{1,5,12}. Poor conduction through QD multilayers¹³ led to an injected charge imbalance, and consequently the luminescence efficiency of these early devices never exceeded 0.10 cd A^{–1}. Furthermore, a high density of pinhole defects in QD multilayers resulted in low device yields and inconsistent device performance. These technological shortcomings are avoided in structures that use only a single monolayer of QDs as the emissive layer. Indeed, high device yields and consistent LED performance are standard for devices described here.

Our basic device structures¹⁴ are shown in Fig. 1 inset, along with a schematic drawing of a core-shell-type QD passivated with trioctylphosphine oxide (TOPO) and trioctylphosphine (TOP) caps. The QD monolayer and the hole-transporting N,N'-diphenyl-N,N'-bis(3-methylphenyl)-(1,1'-biphenyl)-4,4'-diamine (TPD) layer are deposited in a single spin-casting step. Although spin-casting is possible for molecular organics, and typical for organic polymers, it limits the viable organic hole-transporting materials to those that are highly soluble in solvents such as toluene, alkanes and chloroform, which are preferred for solvating TOPO/TOP-capped QD colloids. It is desirable to have organic solubility of the order of 10 mg ml^{–1} to allow for a large range of possible solution mixtures and film thicknesses. Additionally, it is necessary to use an organic hole-transporting material that phase-segregates from the TOPO/TOP-capped QDs during the spinning process. TPD in chloroform meets all of these criteria. The luminescence of TPD also has a significant spectral overlap with the absorption spectra of CdSe QDs of all sizes, allowing the possibility of efficient exciton energy transfer from TPD to QDs.

The QD concentration in TPD/QD solutions is optimized to form a complete single monolayer on top of the TPD film. This is crucial to the creation of LEDs that are both efficient and monochromatic. Atomic force microscopy (AFM) images in Fig. 2a and b show the result of spinning at 1/5 of the optimal concentration, and confirm that TPD and QDs phase-segregate during the spinning process, yielding a QD coverage of 21%. Figure 2c shows a complete

single QD monolayer like those used in devices I and II. Optical absorption measurements of these bilayers indicate that QDs make up 5–10 vol.% of our 40-nm-thick films, confirming that all of the QDs are accounted for by this self-assembled, phase-segregated monolayer. Further confirmation of the phase segregation process is shown in Fig. 2d, where after thermal evaporation of 5 nm of TPD onto the QD monolayer the TPD forms droplets rather than a planar thin film. Thicker organic overlayers can be planar, as in Fig. 2f, which shows the smooth surface of the 40-nm-thick tris-(8-hydroxyquinoline)aluminium (Alq₃) overlayer in device I.

The electroluminescence spectra of our QD-LEDs are shown in Fig. 1. Spectral peaks at wavelengths of 562 and 400 nm, and the broader shoulder centred at 530 nm, are attributed to emission from QDs, TPD and Alq₃, respectively. The dashed lines in Fig. 1a show the decomposition of the electroluminescence spectra into Alq₃, TPD and QD contributions, which we quantify in terms of an emission fraction (f_x).

The external quantum efficiency of device I exceeds $\eta = 0.4\%$ for a broad range of device luminances (from 5 to 2,000 cd m^{–2}), peaking at $\eta = 0.52\%$ at 10 mA cm^{–2} (Fig. 3). The brightness of 100 cd m^{–2} is achieved at current density $J = 5.3$ mA cm^{–2}, voltage $V = 6.1$ V, corresponding to a luminescence efficiency of 1.9 cd A^{–1}. At 125 mA cm^{–2}, the brightness of device I is 2,000 cd m^{–2}, which corresponds to 1.6 cd A^{–1}, and a 25-fold improvement over the best previously reported QD-LED result⁵.

In device I, holes are injected from the indium tin oxide (ITO) contact into the TPD host matrix, and are transported towards the single QD monolayer. Similarly, electrons are injected from the Mg:Ag cathode into the Alq₃ and are transported to the QDs. Exciton generation on QDs occurs via two parallel processes: direct charge injection and exciton energy transfer from organic molecules. For direct charge injection, electrons may be trapped at the QDs owing to the relative energy alignment of the lowest unoccupied molecular orbital (LUMO) levels of TPD, Alq₃ and the QDs (see the energy diagram in Fig. 4). For these charged QDs the barrier

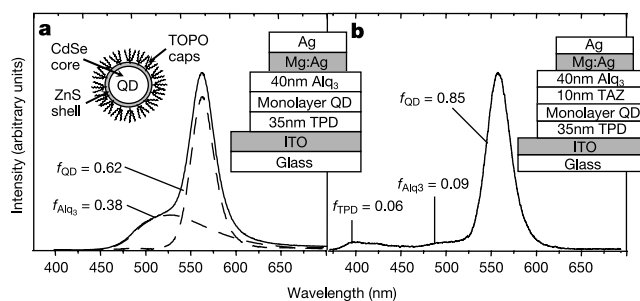


Figure 1 Electroluminescence spectra and structures for two QD-LEDs, devices I and II. Dashed lines, decomposition of spectra into Alq₃ (tris-(8-hydroxyquinoline)aluminium) and QD components. Inset, cartoon of a QD of the core-shell type. Solutions of coated QDs were prepared by the synthetic techniques of refs 4 and 7 with a photoluminescence efficiency of $22 \pm 2\%$. Absorption and luminescence spectra of our QD solutions peak at wavelengths of 545 and 561 nm, respectively, corresponding to a CdSe core diameter of ~ 38 Å coated with 1.5 monolayers of ZnS. The QDs are mixed into a chloroform solution of N,N'-diphenyl-N,N'-bis(3-methylphenyl)-(1,1'-biphenyl)-4,4'-diamine (TPD), which is then spin-cast onto clean, indium tin oxide (ITO) coated glass substrates. The QD and TPD concentrations are optimized such that spin-casting results in the formation of a single QD monolayer on top of a 35-nm-thick TPD layer. **a**, For device I, a 40-nm-thick film of Alq₃ is then thermally evaporated, followed by a 1-mm-diameter, 75-nm-thick Mg:Ag (10:1 by mass) cathode with a 50-nm Ag cap. **b**, For device II, a layer of 3-(4-biphenyl)-4-phenyl-5-*t*-butylphenyl-1,2,4-triazole (TAZ) with a nominal thickness of 10 nm is evaporated before a 30-nm Alq₃ deposition. The spin-casting and device manipulation during growth is performed in a dry nitrogen environment, with moisture and oxygen content of less than 5 p.p.m. All measurements are done in air. For optimized growth conditions device yields are high, comparable to that of thermally evaporated all-organic LEDs. f_x , emission fraction.

to hole injection from the TPD is reduced. Upon acceptance of holes from TPD, excitons form on the QDs, and can subsequently recombine radiatively. Alternatively, excitons can be formed on organic molecules that are near grain boundaries, interstitial spaces, and voids in the single QD monolayer. These excitons can then undergo Förster energy transfer to the lower-energy QD sites, where they recombine radiatively.

Although we observed Förster energy transfer in photoluminescence studies of TPD/QD films, the current–voltage data of our QD-LED (Fig. 3 inset) also suggests generation of excitons on QDs via direct charge injection. Comparison of the current–voltage characteristics of device I with that of a control structure, without the monolayer of QDs, shows that QD-LEDs operate at a consistently higher voltage. The QD monolayer, therefore, inhibits charge conduction, which is consistent with charge-trapping on QD sites. For devices with a QD layer thicker than a single monolayer, the operating voltage increases further while the quantum efficiency is markedly reduced (not shown). This is consistent with the low efficiency and the high operating voltage measured in earlier studies that examined QD-LEDs containing more than ten layers of QDs. Indeed, studies of conduction through close-packed QD layers implicate charge-trapping as an important component of QD conduction^{13,15}. In addition, devices that have only a partial monolayer of QDs have emission spectra that are dominated by Alq₃ (device I) or TPD (device II), indicating that precise control of the assembly of a QD monolayer is critical for optimal device operation.

Device II includes 10 nm of a material that blocks holes and excitons, 3-(4-biphenyl)-4-phenyl-5-*t*-butylphenyl-1,2,4-triazole (TAZ), under the Alq₃ layer. The morphology of this layer, similarly guided by phase segregation (Fig. 2e), results in an incomplete coverage of 72%. The narrow emission spectrum in this device is due to the QDs (FWHM, 32 nm) with $f_{\text{QD}} = 0.85$. The device regions under TAZ voids have a similar structure to device I, and contribute to the observed Alq₃ emission ($f_{\text{Alq}_3} = 0.09$). Elsewhere, TAZ suppresses Alq₃ emission by blocking transport of both holes and excitons into the Alq₃. The minimal TPD emission ($f_{\text{TPD}} = 0.06$) from this device is probably due to incomplete

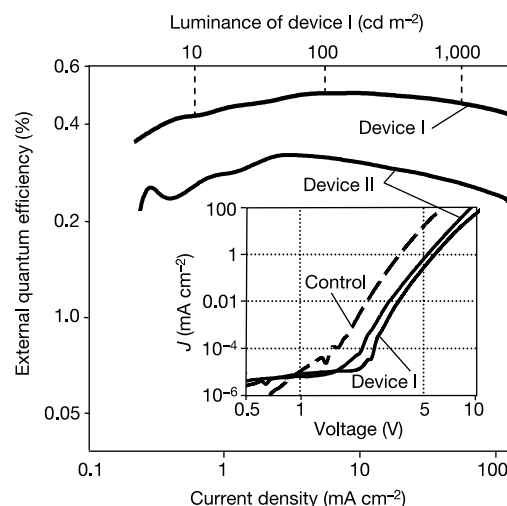


Figure 3 External quantum efficiency versus current density for the two devices shown in Fig. 1. The peak efficiency of device I is $\eta = 0.52\%$ at 10 mA cm^{-2} and 6.6 V , which corresponds to a luminance of 190 cd m^{-2} . Inset, current density versus voltage data for the same devices, as well as for a control device that consists of spin-cast TPD and vacuum-deposited Alq₃ of the same thicknesses as used in device I. In all cases $J \propto V^9$ for $V > 3 \text{ V}$, whereas the devices that contain QDs have operating voltages that are 15–30% greater than the control device.

exciton energy transfer from the TPD molecules to the QDs.

The fundamental limits of QD-LED performance are different from those of all-organic LEDs. The discrete energy structure of QDs gives rise to a narrow emission FWHM, which in our electroluminescence experiments is as small as 32 nm. In contrast, molecular organic LEDs have a typical FWHM of between 50 and 100 nm, although emission of some polymers and phosphorescent molecules was shown to be as narrow as 26 nm FWHM^{16,17}, but with long, low-energy tails. The vibrational structure of structurally flexible organics typically generates broad, single-molecule emis-

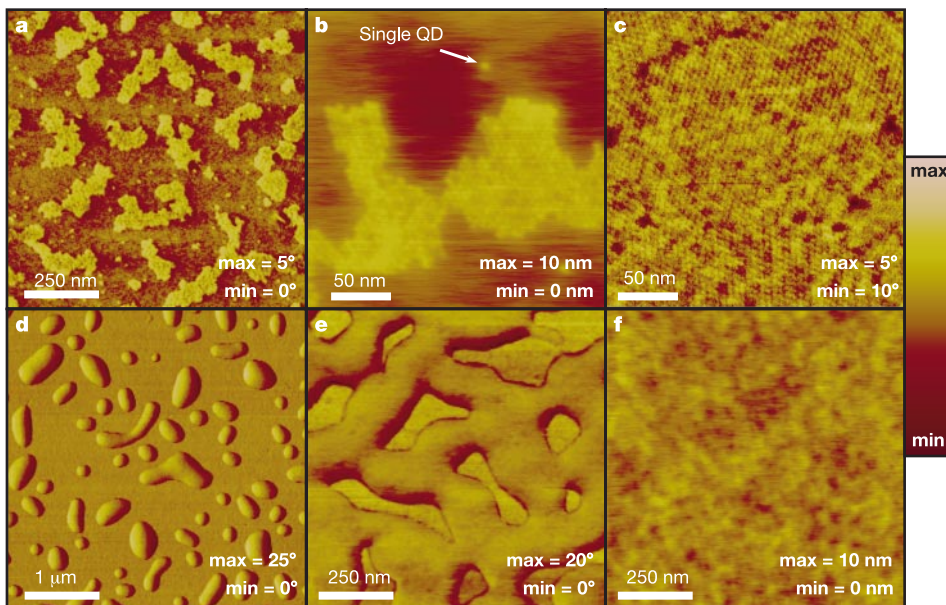


Figure 2 AFM images showing the surface morphology of various organic/QD films. **a**, Phase image of a partial monolayer of QDs on TPD after phase segregation during spin-coating. QD surface coverage is 21%. **b**, Height image of a close-up of **a**, showing both an island of QDs as well as individual QDs on a flat TPD background. **c**, Phase image of a complete, hexagonally packed monolayer of QDs phase segregated from the underlying TPD. Grain boundaries between ordered domains of

QDs are observable. **d**, Phase image of TPD and QDs after thermal evaporation of a nominal 5 nm of TPD onto the complete QD layer of **c**. Material phase segregation results in TPD droplet formation. **e**, Phase image of nominal 10 nm of TAZ on surface similar to **c**. TAZ surface coverage is 72%. **f**, Height image of nominal 40 nm of Alq₃ on surface similar to **c**. Coverage is complete, with an r.m.s. surface roughness of 0.56 nm.

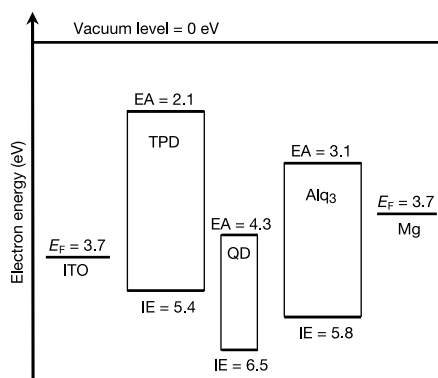


Figure 4 Proposed energy level diagram of device I. Values for ionization energy (IE) of all materials except QDs are taken from photoemission spectroscopy measurements²¹. Electron affinity levels (EA) were then calculated using optical absorption data. QD levels shown are from calculated values¹². E_F , Fermi energy.

sion spectra at room temperature¹⁸. The same is not true of the rigid, covalently bonded inorganic QD, for which single-QD spectroscopy shows that the fundamental FWHM line width at room temperature is 14 nm (ref. 19) with a symmetrical, gaussian shape. It is the combination of spectral diffusion and size distribution of QDs in a sample that yields further line broadening. However, it is reasonable to expect that present-day techniques in QD preparation and processing could yield QD-LED line widths that are as narrow as 25 nm—which has already been accomplished in solution²⁰. Such true colour saturation would benefit applications where efficient production of narrow-band light is desired. In particular, the creation of high-luminous-efficiency red and blue LEDs requires both high external quantum efficiency as well as narrow-band emission, to prevent the bulk of emission from occurring in the infrared or ultraviolet, respectively, where our eyes have minimal response.

We also note that the excited state of QDs mix the spin-triplet and spin-singlet exciton characteristics. In other material systems, the mixed states facilitate capture of both singlet and triplet excitons followed by rapid exciton recombination, leading to demonstrations of LEDs with nearly 100% internal quantum efficiency⁹. It remains an open question as to whether or not analogously efficient exciton harvesting is possible in QD-LEDs.

The present work shows that large area sheets of single QD monolayers ($> \text{cm}^2$) can be used in electrically active devices, minimizing QD material use to only the active region of the device. The phase segregation that governs formation of the organic/QD spin-cast thin film bilayers is a general fabrication process that we expect will be widely applicable. The process is governed by the physical size and chemical character of the two solvated constituents; the TPD molecules are small (~ 1 nm) and have aromatic character, whereas the QDs are comparatively large (> 3 nm) and present a surface that consists of mostly alkane chains. In general, phase segregation is not limited to aromatic/aliphatic pairs, but governs the interaction between any pair of materials with disparate chemical functionality.

Although we have demonstrated a 25-fold improvement in the luminescent power efficiency over previously reported QD-LEDs, we believe that we have not reached the fundamental limits of device performance in terms of both quantum efficiency and colour saturation. □

Received 22 July; accepted 14 October 2002; doi:10.1038/nature01217.

- Kagan, C. R., Mitzi, D. B. & Dimitrakopoulos, C. D. Organic-inorganic hybrid materials as semiconducting channels in thin-film field-effect transistors. *Science* **286**, 945–947 (1999).
- Huynh, W. U., Dittmer, J. J. & Alivisatos, A. P. Hybrid nanorod-polymer solar cells. *Science* **295**, 2425–2427 (2002).
- Colvin, V., Schlamp, M. & Alivisatos, A. Light-emitting diodes made from cadmium selenide nanocrystals and a semiconducting polymer. *Nature* **370**, 354–357 (1994).

- Murray, C. B., Norris, D. J. & Bawendi, M. G. Synthesis and characterization of nearly monodisperse CdE (E = S, Se, Te) semiconductor nanocrystallites. *J. Am. Chem. Soc.* **115**, 8706–8715 (1993).
- Schlamp, M. C., Peng, X. G. & Alivisatos, A. P. Improved efficiencies in light emitting diodes made with CdSe(CdS) core/shell type nanocrystals and a semiconducting polymer. *J. Appl. Phys.* **82**, 5837–5842 (1997).
- Dabbousi, B. O., Bawendi, M. G., Onitsuka, O. & Rubner, M. F. Electroluminescence from CdSe quantum-dot/polymer composites. *Appl. Phys. Lett.* **66**, 1316–1318 (1995).
- Dabbousi, B. O. *et al.* (CdSe)ZnS core-shell quantum dots: Synthesis and characterization of a size series of highly luminescent nanocrystallites. *J. Phys. Chem. B* **101**, 9463–9475 (1997).
- Hines, M. A. & Guyot-Sionnest, P. Synthesis and characterization of strongly luminescent ZnS-capped CdSe nanocrystals. *J. Phys. Chem.* **100**, 468–471 (1996).
- Adachi, C. *et al.* Nearly 100% internal phosphorescence efficiency in an organic light-emitting device. *J. Appl. Phys.* **90**, 5048–5051 (2001).
- Moller, S. & Forrester, S. R. Improved light out-coupling in organic light emitting diodes employing ordered microlens arrays. *J. Appl. Phys.* **91**, 3324–3327 (2002).
- Dirr, S., Bohler, A., Wiese, S., Johannes, H. H. & Kowalsky, W. Organic light emitting diodes with reduced spectral and spatial halfwidths. *Jpn J. Appl. Phys.* **37**, 1457–1461 (1998).
- Mattoussi, H. *et al.* Electroluminescence from heterostructures of poly(phenylene vinylene) and inorganic CdSe nanocrystals. *J. Appl. Phys.* **83**, 7965–7974 (1998).
- Leatherdale, C. A. *et al.* Photoconductivity in CdSe quantum dot solids. *Phys. Rev. B* **62**, 2669–2680 (2000).
- Era, M., Adachi, C., Tsutsui, T. & Saito, S. Double-heterostructure electroluminescent device with cyanine-dye bimolecular layer as an emitter. *Chem. Phys. Lett.* **178**, 488–490 (1991).
- Morgan, N. Y. *et al.* Electronic transport in films of colloidal CdSe nanocrystals. *Phys. Rev. B* **66**, 075339 (2002).
- Liu, J., Shi, Y. J. & Yang, Y. Improving the performance of polymer light-emitting diodes using polymer solid solutions. *Appl. Phys. Lett.* **79**, 578–580 (2001).
- Kwong, R. C. *et al.* Efficient, saturated red organic light emitting devices based on phosphorescent platinum(II) porphyrins. *Chem. Mater.* **11**, 3709–3713 (1999).
- Tamarat, P., Maali, A., Lounis, B. & Orrit, M. Ten years of single-molecule spectroscopy. *J. Phys. Chem. A* **104**, 1–16 (2000).
- Empedocles, S. A., Norris, D. J. & Bawendi, M. G. Photoluminescence spectroscopy of single CdSe nanocrystallite quantum dots. *Phys. Rev. Lett.* **77**, 3873–3876 (1996).
- Murray, C. B., Kagan, C. R. & Bawendi, M. G. Synthesis and characterization of monodisperse nanocrystals and close-packed nanocrystal assemblies. *Annu. Rev. Mater. Sci.* **30**, 545–610 (2000).
- Hill, I. G. & Kahn, A. Organic semiconductor heterointerfaces containing bathocuproine. *J. Appl. Phys.* **86**, 4515–4519 (1999).

Acknowledgements We thank D. Mascaro and E. Shaw for assistance in acquiring AFM images. This work was supported in part by the NSF-MRSEC programme, DMR, and Universal Display Corporation; it made use of MRSEC shared facilities supported by the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to V.B. (e-mail: bulovic@mit.edu).

Metastable garnet in oceanic crust at the top of the lower mantle

Tomoaki Kubo*, Eiji Ohtani*, Tadashi Kondo*, Takumi Kato†, Motomasa Toma*, Tomofumi Hosoya*, Asami Sano*, Takumi Kikegawa‡ & Toshiro Nagase§

* Institute of Mineralogy, Petrology and Economic Geology, Graduate School of Science, Tohoku University, Sendai 980-8578, Japan

† Institute of Geoscience, University of Tsukuba, Tsukuba 305-8571, Japan

‡ Photon Factory, High Energy Accelerator Research Organization, Tsukuba 305-0801, Japan

§ The Tohoku University Museum, Sendai 980-8578, Japan

As oceanic tectonic plates descend into the Earth's lower mantle, garnet (in the basaltic crust) and silicate spinel (in the underlying peridotite layer) each decompose to form silicate perovskite—the 'post-garnet' and 'post-spinel' transformations, respectively. Recent phase equilibrium studies^{1,2} have shown that the post-garnet transformation occurs in the shallow lower mantle in a cold slab, rather than at ~ 800 km depth as earlier studies indicated^{3–6}, with the implication that the subducted basaltic crust is unlikely to become buoyant enough to delaminate as it

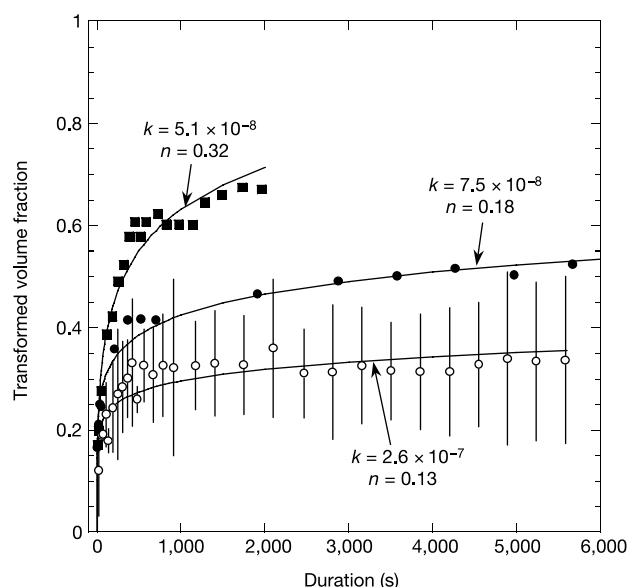


Figure 1 The variation of the transformed volume fraction with duration at the desired temperature. Filled circles and filled squares represent the post-garnet transformation in pure pyrope $\text{Mg}_3\text{Al}_2\text{Si}_3\text{O}_{12}$ (grain size $3.2\ \mu\text{m}$) at 30.8 GPa and 1,273 K ($\Delta P = 5.5\ \text{GPa}$), and at 31.4 GPa and 1,473 K ($\Delta P = 6.0\ \text{GPa}$), respectively. Open circles represent the post-garnet transformation in natural pyrope ($\text{Mg}_{0.724}\text{Fe}_{0.184}\text{Ca}_{0.111}\text{Al}_{0.872}\text{Cr}_{0.044}\text{Ti}_{0.010}\text{Si}_{3.064}\text{O}_{12}$ (grain size $12\ \mu\text{m}$) at 28.1 GPa and 1,600 K (overpressure is 3.5 GPa). Curves and kinetic parameters obtained by least-squares fits of equation (1) are also shown.

enters the lower mantle. But here we report results of a kinetic study of the post-garnet transformation, obtained from *in situ* X-ray observations using sintered diamond anvils, which show that the kinetics of the post-garnet transformation are significantly slower than for the post-spinel transformation⁷. Although metastable spinel quickly breaks down at a temperature of 1,000 K, we estimate that metastable garnet should survive of the order of 10 Myr even at 1,600 K. Accordingly, the expectation of where the subducted oceanic crust would be buoyant spans a much wider depth range at the top of the lower mantle, when transformation kinetics are taken into account.

In situ X-ray diffraction experiments were conducted at the synchrotron radiation facility of Photon Factory (BL14C2, KEK-PF), Tsukuba, Japan. In order to obtain high-quality kinetic data on the post-garnet transformation at pressures of more than 30 GPa, we used Kawai-type high-pressure apparatus⁸ (MAX-III) combined with sintered diamond anvils as the second-stage anvils^{9–12}. Two kinds of pyrope with different chemical compositions were used as the starting material. One is synthetic pure pyrope $\text{Mg}_3\text{Al}_2\text{Si}_3\text{O}_{12}$; the other is natural pyrope in garnet lherzolite from Czechoslovakia, whose composition is $(\text{Mg}_{0.724}\text{Fe}_{0.184}\text{Ca}_{0.111})_3(\text{Al}_{0.872}\text{Cr}_{0.044}\text{Ti}_{0.010})_2\text{Si}_{3.064}\text{O}_{12}$. These garnets decompose at around 25–30 GPa into magnesian-silicate perovskite and corundum¹³, and magnesian-silicate perovskite, calcium-silicate perovskite, aluminous phase and stishovite^{14,15}, respectively. We observed the post-garnet transformation kinetics at 28.1–31.4 GPa and 1,273–1,600 K, where the overpressure from the equilibrium boundary was 3.5–6.0 GPa. Figure 1 shows changes of the transformed volume fraction with time. It was observed that the rate of the post-garnet transformation significantly decreased with time in both starting materials.

Scanning electron microscopy of the recovered samples revealed that the post-garnet transformation in both pyropes occurred at grain boundaries of the parental garnet, and the reaction rim of the decomposed assemblages was formed along the grain boundaries. These observations are consistent with results on the post-spinel transformation at large overpressures (3.5–5.0 GPa; ref. 7). Micro-

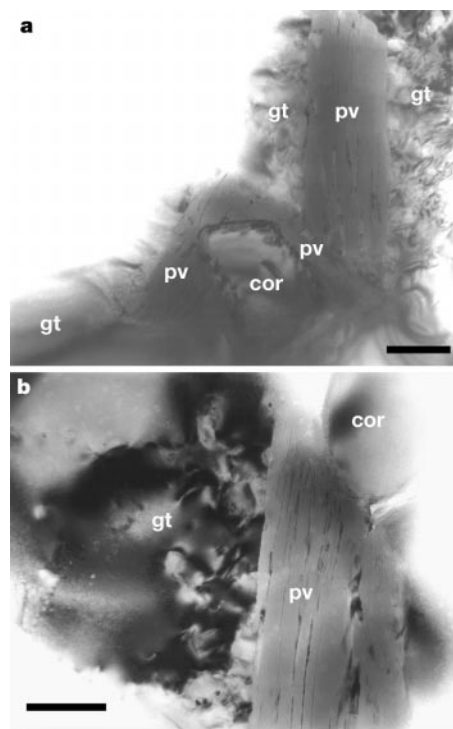


Figure 2 Transmission electron micrographs showing microstructures of the reaction zone in the post-garnet transformation of pure pyrope at 30.8 GPa and 1,273 K. Abbreviations: gt, garnet (that is, pyrope); pv, perovskite; cor, corundum. Scale bars, 200 nm. **a**, Perovskite grows with a rectangular shape, whereas corundum grows with a granular shape. Post-garnet assemblages do not grow with lamellar textures, in contrast to the post-spinel assemblages^{16,17}. **b**, Perovskite often contains fine lamellae of LiNbO_3 structure, which is thought to be formed during decompression¹⁵.

structures of decomposed assemblages in the experiment with pure pyrope were examined using transmission electron microscopy (Fig. 2). We found that the decomposed post-garnet assemblages do not show lamellar growth texture. This contrasts strongly with the post-spinel transformation, in which the decomposed assemblages grow with very fine lamellar texture^{16,17}. These observations suggest that the growth in the post-garnet transformation requires longer-range diffusion than that in the post-spinel transformation.

We analysed the present kinetic data on the basis of the observed transformation mechanisms. The overall transformation rate is thought to be controlled only by the growth of the reaction rim, because the rim was formed in the initial stage of the transformation owing to the high nucleation rate on grain boundaries at large overpressure conditions. In this case, the rate equation is given by¹⁸

$$V = 1 - \exp[-2SX(t)] \quad (1)$$

where V is the transformed volume fraction, S is the area of the grain boundary, and $X(t)$ is the growth distance at time t . The area of grain boundary can be expressed as $3.35/d$, where d is the grain size of the parental phase. $X(t)$ is described by kt^n , where k and n are constants. We fitted this rate equation to the kinetic data using a nonlinear least-squares procedure to optimize kinetic parameters k and n (Fig. 1). The value of n was estimated to be very small and less than 1 in the post-garnet transformation, which means that the growth rate is time dependent and significantly decreases with time. This is different from the post-spinel transformation, in which n was nearly 1 and the growth rate is almost constant at large overpressure conditions⁷.

These differences in growth kinetics are believed to originate from differences in growth textures between these transformations. In case of the post-spinel transformation, the diffusion distance

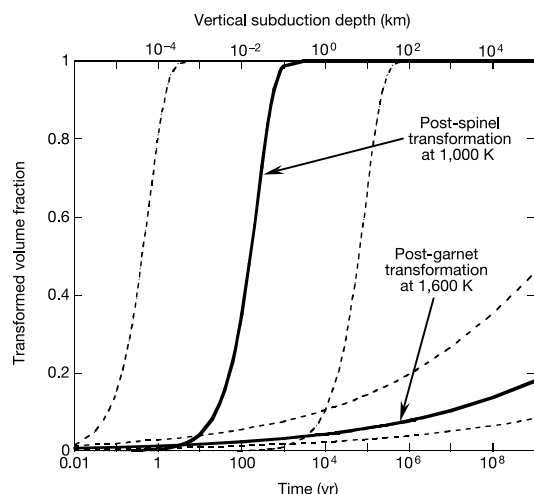


Figure 3 Comparison between the rates of the post-spinel and post-garnet transformations under subduction-zone conditions. We assume a vertical subduction velocity of 10 cm yr^{-1} for the slab. The post-spinel transformation rate was estimated at 1,000 K, an overpressure of 1 GPa, and a grain size of 1 mm, using kinetic parameters from ref. 7. The post-garnet transformation rate was estimated at 1,600 K, an overpressure of 3.5 GPa, and a grain size of 1 mm, using the kinetic parameters of natural pyrope determined in this study. Dashed lines represent errors in the transformation rate.

needed for the decomposition is constant, owing to lamellar growth at constant lamellar spacing, resulting in the constant growth rate^{7,17}. On the other hand, in case of the post-garnet transformation, the diffusion distance needed for the decomposition becomes larger as the decomposed assemblages grow, resulting in the slow-down of the growth rate (that is, diffusion-controlled growth). Consequently, the rate of the post-garnet transformation is much slower than that of the post-spinel transformation.

We now estimate the rate of these compositional transformations under cold subduction zone conditions, assuming that the overall transformation rate is controlled only by the growth (Fig. 3). Kinetic parameters estimated in the previous study⁷ were used for the post-spinel transformation. For the post-garnet transformation, the kinetic parameters for natural pyrope estimated in the present experiment were used; this is because natural pyrope is likely to be closer in composition to subducted oceanic crust than is pure pyrope. The temperature and the grain size of the parental phase in the slab will also affect the transformation kinetics. Temperatures in the cold subducting slab near the 660-km discontinuity are estimated to be 1,400–1,600 K at the top of the slab (the basaltic crust) and about 1,000 K at the centre of the slab (the peridotite layer)¹⁹. The grain size of spinel in the cold slab is believed to be less than 1 mm owing to the olivine–spinel transformation²⁰. But the grain size of garnet is probably greater than 1 mm because the garnet in natural eclogite is relatively large (of the order of 1 cm) and it undergoes no transformations above the lower mantle. Figure 3 shows that the post-spinel transformation completes in $\sim 10^3$ yr even at the relatively low temperature of 1,000 K and with a grain size of 1 mm. Although nucleation kinetics may control the depth of the post-spinel transformation, significant metastable spinel does not exist in the peridotite layer of the cold slab. In contrast, the post-garnet transformation does not complete in geological timescales of $\sim 10^6$ – 10^7 yr, even at a higher temperature of 1,600 K and a grain size of 1 mm. Therefore, a large amount of metastable garnet could exist in the cold descending oceanic crust into the lower mantle.

We plot changes in the density of the subducted oceanic crust (mid-ocean-ridge basalt) with depth due to the post-garnet transformation in both equilibrium²¹ and metastable models in Fig. 4. In the equilibrium model at 1,700 K, the post-garnet transformation occurs between 600 and 800 km depth, and the subducted oceanic

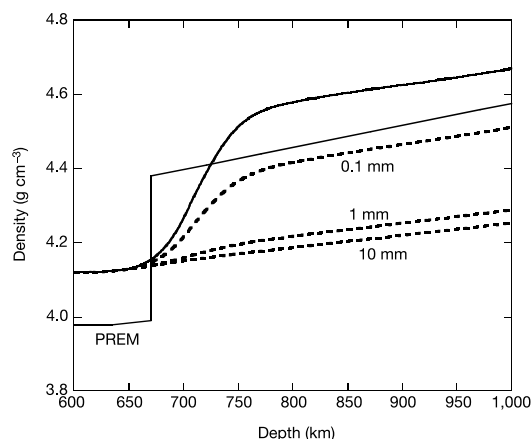


Figure 4 Effects of transformation kinetics on the density of oceanic crust (mid-ocean-ridge basalt, MORB) descending into the lower mantle. The solid curve represents the equilibrium model of MORB density at 1,700 K (ref. 21). The dashed curves represent the metastable model of MORB density at 1,600 K with grain sizes of 0.1, 1 and 10 mm, taking the transformation kinetics into account. The average mantle density²⁸ (using the PREM Earth model) is also shown (thin straight lines).

crust becomes denser than the surrounding mantle at around 720 km depth. The density change in the metastable model was calculated from multiplying that in the equilibrium model by the transformed fraction estimated in Fig. 3. It was assumed that the transformation starts at 600 km depth and that the vertical subduction velocity is 10 cm yr^{-1} . Figure 4 suggests that when the temperature is less than 1,600 K, the subducted oceanic crust is buoyant to depths greater than $\sim 1,000$ km, taking into account the kinetics of the post-garnet transformation.

Because the phase boundaries of the post-spinel and post-garnet transformations have negative and positive slopes, respectively, the depth of these transformations in the cold slab becomes closer in the equilibrium model^{1,2}. So this model predicts that the subducted oceanic crust could sink into the lower mantle without the formation of the garnetite layer. However, the present study demonstrates that differences in kinetics between these transformations have an important effect on the dynamics of the cold slab descending into the lower mantle. A buoyancy force against subduction, due to the presence of the metastable garnet, would enhance the separation of the thin oceanic crust from the peridotite layer and the accumulation of the metastable garnetite at the top of the lower mantle. It has been shown that the separation of the oceanic crust is possible from the mineral-physics point of view²². The trapped metastable garnetite may be responsible for seismic discontinuities at depths of 900–1,080 km observed beneath subduction zones^{23,24}.

When the metastable garnetite is heated, so that the transformation can proceed, it is no longer buoyant and the slab sinks into the deep mantle. Furthermore, the basaltic portion of the cold slab may contain a small amount of water in nominally anhydrous phases²⁵, which might affect the post-garnet transformation kinetics and the fate of the metastable garnetite. Further studies on the thermal evolution of the slab, and on the effects of temperature and water on transformation kinetics, could tell us how long the metastable garnetite survives, and how much remains at the top of the lower mantle. □

Methods

The sample assembly is composed of a sintered magnesia pressure medium, a cylindrical Re or LaCrO₃ heater, a molybdenum electrode and a graphite sample capsule. The edge length of the second-stage sintered diamond anvil was 10 or 14 mm. In the latter system, one advanced diamond composite (ADC)^{12,26} of 14 mm edge length (manufactured by ANUTECH) was used in order to increase X-ray transmissivity. The truncated edge length of the second-stage anvil is 2.0 mm. Pressure was calculated from the unit cell volume of gold²⁷, and the temperature was monitored by a W3%Re–W25%Re thermocouple. Equilibrium boundaries of the present post-garnet transformations have been determined

by *in situ* X-ray observations based on the same pressure scale in the previous study^{13,14}. The overpressure was calculated using these boundaries. We observed the pressure to drop during the transformation by 1–2 GPa, but this would not affect the transformation rate significantly because the overpressure is very large in the present study. This has been confirmed in the post-spinel transformation⁷.

One starting material was a powdered mixture of natural pyrope and gold. In experiments using the natural pyrope, the sample was annealed at 20 GPa and 1,523 K for 2 h before the transformation to achieve equilibrium microstructures, resulting in equigranular polycrystalline pyrope of $12 \pm 2 \mu\text{m}$ in diameter. Another starting material was a sintered mixture of pure pyrope and gold, which was synthesized at 20 GPa and 1,773 K using the pyrope glass. The sintered pure pyrope was not annealed before the transformation because equilibrium microstructures in the sintered sample are almost maintained during the cold compression stage⁷. The grain size of pure pyrope was $3.2 \pm 0.5 \mu\text{m}$. After the pressure reached the desired value, the sample was heated at a rate of 500 K min^{-1} . When the temperature reached the desired value, it was kept constant and X-ray diffraction patterns of the sample were taken every 10–300 s by the energy dispersive method using a Ge solid-state detector. The transformed volume fraction was estimated on the basis of the integrated intensity of diffraction lines 400, 642 in pyrope relative to the intensity before the transformation⁷. We observed uniform changes of the diffraction peaks during the transformation, which is diagnostic of the absence of a preferred orientation.

Received 29 July; accepted 25 October 2002; doi:10.1038/nature01281.

- Hirose, K., Fei, Y., Ma, Y. & Mao, H. K. The fate of subducted basaltic crust in the Earth's lower mantle. *Nature* **397**, 53–56 (1999).
- Akaogi, M. & Ito, E. Calorimetric study on majorite-perovskite transition in the system $\text{Mg}_3\text{Si}_2\text{O}_{12}$ – $\text{Mg}_3\text{Al}_2\text{Si}_2\text{O}_{12}$: transition boundaries with positive pressure-temperature slopes. *Phys. Earth Planet. Inter.* **114**, 129–140 (1999).
- Irifune, T. & Ringwood, A. E. Phase transformations in subducted oceanic crust and buoyancy relationships at depths of 600–800 km in the mantle. *Earth Planet. Sci. Lett.* **117**, 101–110 (1993).
- Ringwood, A. E. Role of the transition zone and 660 km discontinuity in mantle dynamics. *Phys. Earth Planet. Inter.* **86**, 5–24 (1994).
- Kesson, S. E., Fitz Gerald, J. D. & Shelley, J. M. G. Mineral chemistry and density of subducted basaltic crust at lower-mantle pressures. *Nature* **372**, 767–769 (1994).
- Faust, J. & Knittle, E. The stability and equation of state of majoritic garnet synthesized from natural basalt at mantle conditions. *Geophys. Res. Lett.* **23**, 3377–3380 (1996).
- Kubo, T. *et al.* Mechanisms and kinetics of the post-spinel transformation in Mg_2SiO_4 . *Phys. Earth Planet. Inter.* **129**, 153–171 (2002).
- Kawai, N. & Endo, S. The generation of ultrahigh hydrostatic pressures by a split sphere apparatus. *Rev. Sci. Instrum.* **41**, 1178–1181 (1970).
- Ohtani, E. *et al.* High pressure generation by a multiple anvil system with sintered diamond anvils. *Rev. Sci. Instrum.* **60**, 922–925 (1989).
- Kondo, T. *et al.* Ultrahigh-pressure and high-temperature generation by use of the MA-8 system with sintered-diamond anvils. *High Temp. High Press.* **25**, 105–112 (1993).
- Kato, T., Ohtani, E., Kamaya, N., Shimomura, O. & Kikegawa, T. *High Pressure Research in Mineral Physics: Application to Earth and Planetary Sciences* (eds Syono, Y. & Manghnani, M. H.) 33–36 (Geophysical Monograph 67, American Geophysical Union, Washington DC, 1992).
- Kato, T. *et al.* *In situ* X-ray observation of high-pressure phase transitions of MgSiO_3 and thermal expansion of MgSiO_3 perovskite at 25 GPa by double-stage multianvil system. *J. Geophys. Res.* **100**, 20475–20481 (1995).
- Hirose, K., Fei, Y., Ono, S., Yagi, T. & Yagi, T. *In situ* measurements of the phase transition boundary in $\text{Mg}_3\text{Al}_2\text{Si}_2\text{O}_{12}$: implications for the nature of the seismic discontinuities in the Earth's mantle. *Earth Planet. Sci. Lett.* **184**, 567–573 (2001).
- Oguri, K. *et al.* Post-garnet transition in a natural pyrope: a multi-anvil study based on *in situ* X-ray diffraction and transmission electron microscopy. *Phys. Earth Planet. Inter.* **122**, 175–186 (2000).
- Miyajima, N., Fujino, K., Funamori, N., Kondo, T. & Yagi, T. Garnet-perovskite transformation under conditions of the Earth's lower mantle: an analytical transmission electron microscopy. *Phys. Earth Planet. Inter.* **116**, 117–131 (1999).
- Poirier, J. P., Peyronneau, J., Madon, M., Guyot, F. & Revcolevschi, A. Eutectoid phase transformation of olivine and spinel into perovskite and rock salt structures. *Nature* **321**, 603–605 (1986).
- Kubo, T. *et al.* Formation of metastable assemblages and mechanisms of the grain-size reduction in the postspinel transformation of Mg_2SiO_4 . *Geophys. Res. Lett.* **27**, 807–810 (2000).
- Cahn, J. W. The kinetics of grain boundary nucleated reactions. *Acta Metall.* **4**, 449–459 (1956).
- Kirby, S. H., Stein, S., Okal, E. A. & Rubie, D. C. Metastable mantle phase transformations and deep earthquakes in subducting oceanic lithosphere. *Rev. Geophys.* **34**, 261–306 (1996).
- Riedel, M. R. & Karato, S. Grain-size evolution in subducted oceanic lithosphere associated with the olivine-spinel transformation and its effects on rheology. *Earth Planet. Sci. Lett.* **148**, 27–43 (1997).
- Ono, S., Ito, E. & Katsura, T. Mineralogy of subducted basaltic crust (MORB) from 25 to 37 GPa, and chemical heterogeneity of the lower mantle. *Earth Planet. Sci. Lett.* **190**, 57–63 (2001).
- Karato, S. On the separation of crustal component from subducted oceanic lithosphere near the 660 km discontinuity. *Phys. Earth Planet. Inter.* **99**, 103–111 (1997).
- Kawakatsu, H. & Niu, F. Seismic evidence for a 920-km discontinuity in the mantle. *Nature* **371**, 301–305 (1994).
- Niu, F. & Kawakatsu, H. Depth variation of the mid-mantle seismic discontinuity. *Geophys. Res. Lett.* **24**, 429–432 (1997).
- Pawley, A. R., McMillan, P. F. & Holloway, J. R. Hydrogen in stishovite, with implications for mantle water content. *Science* **261**, 1024–1026 (1993).
- Irifune, T. *et al.* *Properties of Earth and Planetary Materials at High Pressure and Temperature* (eds Manghnani, M. H. & Yagi, T.) 1–8 (Geophysical Monograph 101, American Geophysical Union, Washington DC, 1998).
- Anderson, O. L., Issak, D. G. & Yamamoto, S. Anharmonicity and the equation state for gold. *J. Appl. Phys.* **6**, 1534–1543 (1989).
- Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model. *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).

Acknowledgements We thank D. H. Green and B. Hiberson for discussions and information on the larger ADC anvil, and K. Fujino for providing natural pyrope crystal and comments. This work was partially supported by the Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T. Kubo (e-mail: tkubo@mail.cc.tohoku.ac.jp).

SAR11 clade dominates ocean surface bacterioplankton communities

Robert M. Morris*, Michael S. Rappé*, Stephanie A. Connon*, Kevin L. Vergin*, William A. Siebold*, Craig A. Carlson† & Stephen J. Giovannoni*

* Department of Microbiology, Oregon State University, Corvallis, Oregon 97331, USA

† Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106-9610, USA

The most abundant class of bacterial ribosomal RNA genes detected in seawater DNA by gene cloning belongs to SAR11—an α -proteobacterial clade¹. Other than indications of their prevalence in seawater, little is known about these organisms. Here we report quantitative measurements of the cellular abundance of the SAR11 clade in northwestern Sargasso Sea waters to 3,000 m and in Oregon coastal surface waters. On average, the SAR11 clade accounts for a third of the cells present in surface waters and nearly a fifth of the cells present in the mesopelagic zone. In some regions, members of the SAR11 clade represent as much as 50% of the total surface microbial community and 25% of the subeuphotic microbial community. By extrapolation, we estimate that globally there are 2.4×10^{28} SAR11 cells in the oceans, half of which are located in the euphotic zone. Although the biogeochemical role of the SAR11 clade remains uncertain, these data support the conclusion that this microbial group is among the most successful organisms on Earth.

Comprehensive information is available about the overall abundance of marine bacterioplankton, which dominate living biomass and have a chief ecological role in marine food webs^{2–5}. With the exception of marine Archaea^{6,7}, however, there is very little quan-

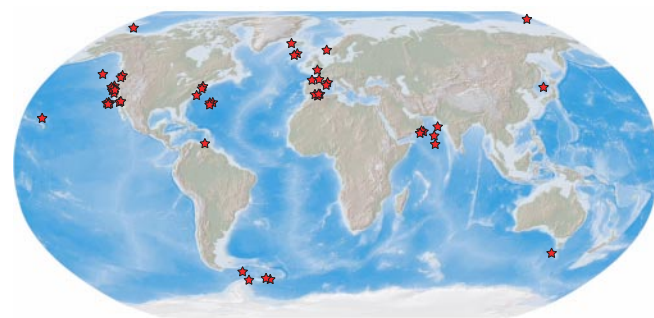


Figure 1 Distribution of the SAR11 clade in the world's oceans. Red marks indicate locations where SAR11 ribosomal RNA genes have been detected using molecular techniques. Although 52 individual samples are represented, many studies along the US Pacific coast and in the North Atlantic have been conducted at the same sample locations and are therefore represented by a single red mark.

titative information about the cellular abundance of specific phylogenetic groups of marine microorganisms. Most knowledge about bacterioplankton diversity comes from 16S ribosomal RNA gene (ribosomal DNA) cloning studies, which have identified several groups of marine bacteria that are thought to be abundant in seawater¹. Although members of the SAR11 clade have no ascribed physiology, they consistently dominate 16S rDNA clone libraries. They have been present in over 50 studies of marine bacterioplankton microbial diversity from sites around the globe (Fig. 1) and account for 25% of all the genes recovered in these studies. Although gene clone library data strongly support the importance of the SAR11 clade in the oceans, these data are open to considerable interpretation, owing to the uncertainties associated with the specificity of polymerase chain reaction (PCR) amplification primers, ribosomal RNA gene copy number and other factors⁸.

In August 2001, seawater samples were collected along a meridional transect in the northwestern Sargasso Sea. Additional samples collected from the Bermuda Atlantic Time-series Study (BATS) site in February and March 2001 were also processed for abundance during winter and spring periods. Surface samples (10 m depth) from the Oregon coast were collected along the Newport Hydroline at station NH5 (44° 39.1' N, 124° 10.6' W) in September 2001 and February 2002, where average nutrient content and productivity are markedly higher than in the western Sargasso Sea. Four Cy3-labelled oligonucleotide probes targeting SAR11 clade 16S rRNA and five Cy3-labelled oligonucleotide probes targeting bacterial 16S rRNA were used in separate hybridization reactions. Digital processing of raw images was used to restrict counting to cells showing fluorescence from both the Cy3 probe and the nucleic acid dye 4',6-diamidino-2-phenylindole dihydrochloride (DAPI)⁹. With this approach, there was very little ambiguity in determining cell counts.

Unlike gene cloning, fluorescence *in situ* hybridization (FISH) is an accurate method for determining the exact abundance of cells in natural samples^{10–12}, provided that the probes circumscribe their phylogenetic target and the fluorescence signals from cells are high enough for consistent scoring. Because some microbial cells, such as small, slow-growing bacterioplankton, are difficult to detect by this method, various strategies have been used to increase signal intensity and counting accuracy. Some of these studies have focused primarily on probe design^{13,14}, whereas others have explored strategies to amplify the fluorescent signal per cell^{15–17}. Our strategy was to use several probes that target different regions of the 16S rRNA to produce an additive effect on signal intensity, coupled with a sensitive cooled CCD (charge-coupled device) camera for detecting

low levels of fluorescence. Our measurements of bacterial abundances were equivalent to those obtained with polynucleotide probes (poly-FISH) in coastal North Sea and Monterey Bay Californian samples¹⁸. The high quantum efficiency of cooled CCD cameras makes it theoretically possible to detect single probe molecules; thus, in hybridization experiments such as these the real limitations are the signal-to-noise ratio achieved and the specificity of the probes.

Analyses of hybridization images showed that cells hybridizing to the SAR11 probes were abundant curved rods of less than 1 μm (Fig. 2). These accounted for 35% ($n = 24$) and 18% ($n = 7$) of total microbial cell abundance in the euphotic and mesopelagic zones, respectively. There was no discernible difference in the morphology of SAR11 cells from the different oceans, or from different depths in the Atlantic. The biovolume of SAR11 cells relative to the average for the marine bacterioplankton community was determined from the ratio of the average areas of SAR11 cell images ($n = 175$) and bacterial cell images ($n = 257$) by assuming that the cells were prolate spheroids¹⁹. The biovolume of SAR11 cells was inferred to be roughly half of the average biovolume for cells in the bacterioplankton community.

Depth profiles showed that members of the SAR11 clade were numerically abundant throughout the water column in the western Atlantic, averaging $2.0 \times 10^8 \text{ cells l}^{-1}$ ($n = 24$) within and $0.2 \times 10^8 \text{ cells l}^{-1}$ ($n = 7$) below the euphotic zone ($>150 \text{ m}$; Fig. 3). Cell counts determined with DAPI averaged $5.7 \times 10^8 \text{ cells l}^{-1}$ ($n = 72$) within and $1.2 \times 10^8 \text{ cells l}^{-1}$ ($n = 21$) below the euphotic zone. To ensure that cells were not being removed in the hybridization process, direct cell counts determined from hybridization preparations were compared with independent preparations that used standard DAPI staining procedures for counting cells²⁰. Agreement between the values was 99.6% in pair-wise comparisons, showing that most cells remained on the filters throughout the hybridization procedure.

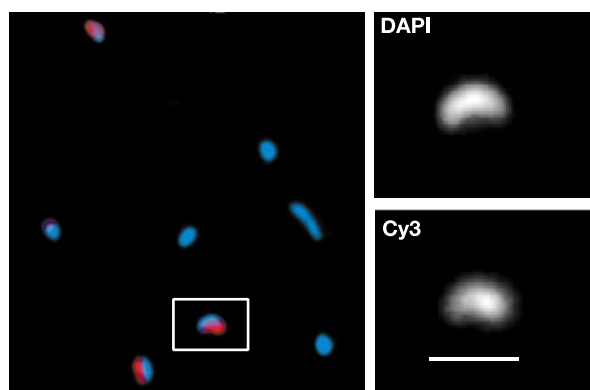


Figure 2 SAR11 fluorescence *in situ* hybridization image composite. Dual image overlay of DNA-containing cells stained with DAPI (blue) and the Cy3 probe (red). Cells emitting a signal for both DAPI and the Cy3 probe are both blue and red, and cells that did not hybridize to the set of SAR11 probes are blue. The identical fields of view in the DAPI- and Cy3-stained images show the characteristic size and curved rod morphology of a magnified SAR11 cell (white box). Scale bar, 1 μm .

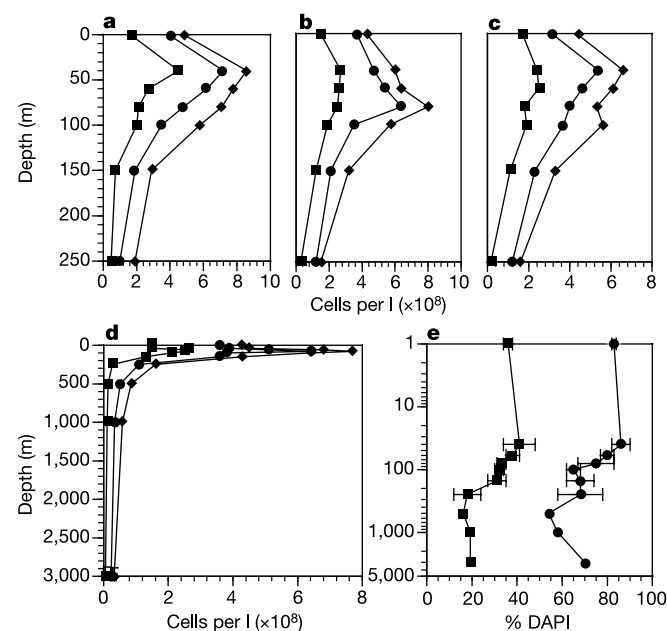


Figure 3 SAR11 probe counts, bacterial probe counts and direct cell counts (DAPI-staining particles) in the northwestern Sargasso Sea. **a–d**, SAR11 clade (squares), Bacteria (circles) and DAPI (diamonds) counts at 32° N, 64° W (CDOM-01; BATS site; **a**), 30° N, 64° W (CDOM-03; **b**), 28° N, 64° W (CDOM-05; **c**) and 26° N, 64° W (CDOM-07; **d**). **e**, A transect composite shows the mean abundance values by depth for SAR11 clade and bacterial cell counts as percentages of direct cell counts (DAPI staining particles); $n = 4$, except for depths below 250 m, where $n = 1$. Standard deviations are given for depths of 1–250 m. One sample point was obtained for depths below 250 m at 26° N, 64° W (CDOM-07).

Probe counts were compared with total DAPI cell counts from the same samples to determine the relative abundances of the SAR11 clade and domain Bacteria. Results from a mean composite of all Atlantic transect profiles indicated that SAR11 accounted for 31–41% of the total DAPI cell counts within the euphotic zone, and 16–19% between 250 and 3,000 m (Fig. 3). The relative abundance of SAR11 exceeded 40% in several profiles and accounted for 51% of the microbial community in the 40-m sample from the BATS site (CDOM-01). Data from samples collected at BATS in February and March 2001 showed similarly high cellular abundances of SAR11 cells in winter months (data not shown). On average, SAR11 cells accounted for 35% ($n = 8$) of the total DAPI cell counts in surface waters to 200 m, and 16% ($n = 5$) at depths from 250 m to 3,600 m. The relative abundance of the SAR11 clade in Oregon coastal surface samples was 17% in September 2001 and 38% in February 2002.

Transect averages showing the relative contribution of cells counted with the bacterial probe suite to total DAPI cell counts ranged from 65 to 86% in the upper ocean surface, reaching maximum values in the surface 40 m (Fig. 3). These values were corrected by subtracting the autofluorescent cell counts obtained from duplicate slides hybridized to the negative control probe. This procedure may result in an underestimate of the total number of cells detected with the bacterial probes, because autofluorescent cells, including cyanobacteria, were excluded. The bacterial contribution to total DAPI cell counts declined to 54% between 250 and 3,000 m. Although bacterial cell counts may be underestimated because some cells fail to produce a probe signal that is sufficient to raise the detection level above background, these observations are consistent with previous findings of a significant archaeal contribution to picoplankton abundance in the mesopelagic⁶.

Hybridization of radioactive probes to bulk RNA extracted from the same samples and from monthly time-series samples collected at the BATS site between September 1997 and August 2000 showed a similar trend in the vertical distribution of SAR11 clade rRNA (Fig. 4). In transect samples taken in August 2001, the percentage of bacterial 16S rRNA contributed by the SAR11 clade varied from 18% ($n = 3$) in the euphotic zone to 3% ($n = 4$) in the aphotic zone (Fig. 4b). Monthly time-series data produced similar values and showed an annual trend in SAR11 clade 16S rRNA abundance (Fig. 4a). Maximal values occurred in the summer, when SAR11 clade 16S rRNA accounted for as much as 32%, 26% and 28% of all bacterial 16S rRNA present in surface waters at the BATS site (June–July 1998, August 1999 and August 2000, respectively). Note that the relatively low temporal variability of total DAPI counts

(Fig. 4c), despite a temporally dynamic SAR11 population, suggests the possibility that other prokaryotic populations are changing as well. SAR11 SSU rRNA percentages are less than the measured percentages of SAR11 cells for the same depth strata, probably because SAR11 cells are unusually small and ribosome content is correlated with cell volume²¹. SAR11 cells have been cultured in our laboratory, and SAR11 biovolumes have been estimated to be about $0.01 \mu\text{m}^3$ from electron micrograph data²². These results and the image data discussed above indicate that the SAR11 contribution to marine biomass is less than their absolute numbers would suggest.

A previous study that used a single oligonucleotide probe to count SAR11 cells in coastal water samples reported that SAR11 abundance was low (<1% of total marine microbial abundance)²³. That result may reflect a real variation in SAR11 abundance; alternatively, single oligonucleotide probes might undercount their target if signals from some cells fall below the background fluorescence of the *in situ* hybridization preparation, which is particularly likely to occur when the target cells are unusually small¹⁸.

Data showing a global distribution of SAR11 (Fig. 1), and the consistent cell counts that we obtained from highly disparate sites, suggest that global extrapolation from our data is justified. Using our mean percentages for surface and mesopelagic samples and previous reported values²⁴, which estimated that globally there are 3.6×10^{28} prokaryotic cells in the upper 200 m of the ocean surface and 6.5×10^{28} prokaryotic cells below 200 m, we calculate that the global abundance of SAR11 in the oceans is 2.4×10^{28} cells. Previous studies⁶ have calculated the abundance of the pelagic crenarchaeota clade to be of the same order as our estimates for the SAR11 clade, 1.3×10^{28} archaeal cells, and have calculated bacterial cell numbers of 3.1×10^{28} cells. On the basis of previous calculations^{6,24} of total global ocean prokaryotic cell abundance, we estimate that SAR11 contributes 24–55% of oceanic prokaryotic cells. Using estimates of global ocean prokaryotic abundance²⁴ and the relative cell sizes measured by microscopic imaging, we calculate that SAR11 cells account for 18% of the total bacterial biomass in surface waters (upper 200 m) and for 9% in deeper waters (200–3,000 m), assuming that the average SAR11 cell contains roughly half as much carbon per cell as the average for marine bacteria. Overall, we estimate that SAR11 contributes about 12% of total marine prokaryotic biomass. Although approximations, these numbers highlight the magnitude of global SAR11 clade populations.

Dividing SAR11 cells were observed in Sargasso Sea samples from all depths, suggesting that members of this clade are actively

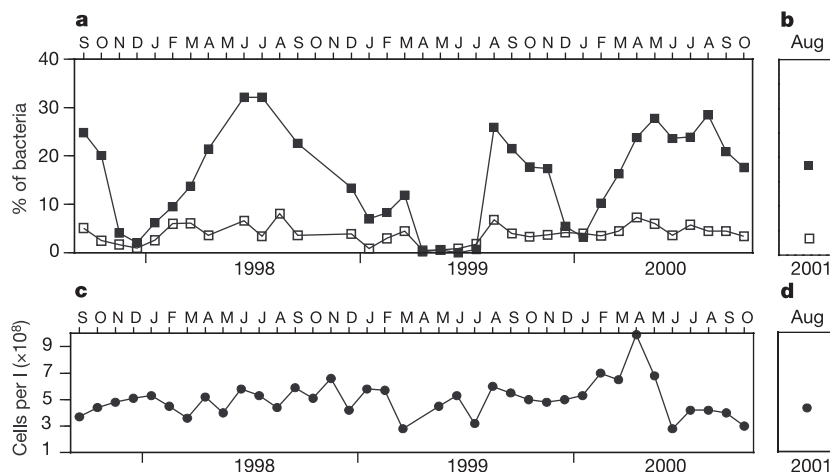


Figure 4 Percentages of SAR11 clade 16S rRNA at surface depths and depths ≥ 200 m. **a**, Monthly time-series samples from the BATS site for surface (filled squares) and 200-m depths (open squares) over a 3-yr period. **b**, August 2001 transect averages for surface ($n = 3$) and ≥ 200 m ($n = 4$) depths from CDOM-03

(1 m), CDOM-05 (1, 250 m) and CDOM-07 (1, 250, 1,000 and 3,000 m).

c, Corresponding total prokaryotic abundance (filled circles) in the surface 5 m for the BATS time series data, estimated by direct counting with DAPI. **d**, Average prokaryotic abundance from 1 m (CDOM-01, CDOM-03, CDOM-05 and CDOM-07).

replicating throughout the water column. Although exponentially growing cultures can be readily distinguished from stationary-phase cultures by the presence of dividing cells in the former, we were unable to make quantitative estimates of the *in situ* growth rates of SAR11 cells by measuring the frequency of dividing cells. The noise in image analysis parameters associated with cell shape is inherently high when viewing cells that have a size close to the limit of resolution of light microscopy.

Although the SAR11 clade probe suite used in this study did not distinguish between different SAR11 lineages, previous work has shown that different phylogenetic groups within the SAR11 clade are found in different regions of the water column²⁵. It is unknown whether this phylogenetic differentiation is associated with specialization in the use of nutrient resources. The numbers of SAR11 cells indicated by this study suggest that these organisms are efficient competitors for resources that are available throughout the water column. Our results do not rule out the possibility that other microorganisms may grow more rapidly than SAR11 but may be less abundant because grazing, viral predation or other sources of mortality result in a high rate of cell removal. Genome sequencing of SAR11 is now in progress, and future research will undoubtedly seek to identify specifically the physiological activities of these cells and their role in the oceanic carbon cycle. □

Methods

Sample collection

We collected water from four transect stations in the northwestern Sargasso Sea. Samples from ten depths between 1 and 3,000 m were collected in Niskin bottles on a 24-place conductivity–temperature–density device rosette and transferred to primary collection bottles. Subsample volumes of 10–40 ml were immediately fixed in filtered paraformaldehyde at a final concentration of 3% and stored at 4 °C for 6–8 h. Fixed samples were filtered onto white 0.2-µm Osmonics polycarbonate filters, placed immediately in slide boxes containing silicon desiccant and stored at –20 °C.

Probe analysis

We analysed 261 SAR11 and 12,300 total 16S rRNA sequences using the ARB sequence analysis package²⁶ to determine probe specificity and accuracy. A total of 149 SAR11 sequences contained sequence spanning at least one target site, and 133 contained an exact match to at least one probe sequence. A minimum of three of the four probes matched exactly all rRNA genes with sequence spanning all target sites ($n = 27$), and most of these sequences (66%) matched all four probes identically. None of the probes used in the SAR11 probe suite matched sequences outside the SAR11 clade. The oligonucleotide probes used to enumerate members of the SAR11 clade were as follows: SAR11-152R (5′-ATTAG CACAAGTTTCCYCGTGT-3′), SAR11-441R (5′-TACAGTCATTTCCTCCCGAC-3′), SAR11-542R (5′-TCCGAAGTACGCTAGGTC-3′), and SAR11-732R (5′-GTCAGTAATG ATCCAGAAAGTYTG-3′). The oligonucleotide probes used to enumerate Bacteria were as follows: EUB-27R (5′-CTGAGCCAKAGTCAACTCT-3′), EUB-338Rpl (5′-GCWGCC WCCCGTAGGWTG-3′), EUB-700R (5′-CTAHGCATTTACACGCTACAC-3′), EUB-700Ral (5′-CTACGAATTCACCTCTACAC-3′) and EUB-1522R (5′-AAGGAGGTGAT CCANCCVCA-3′). The negative control oligonucleotide was 338F (5′-TGAGGATGCCC TCCGTCG-3′).

FISH

Hybridization reactions were carried out essentially as described¹² with the following modifications. Reactions were done on one-quarter membrane sections at 37 °C for 16 h in hybridization buffer (900 mM NaCl, 20 mM Tris (pH 7.4), 0.01% (w/v) SDS, 15% formamide) and either SAR11 or Bacteria-specific Cy3-labelled oligonucleotide probes. Individual probes in each set of probes were each used at a final concentration of 2 ng µl⁻¹. We also used a Cy3-labelled nonsense oligonucleotide (338F) as a negative control at a final concentration of 8 ng µl⁻¹. Optimal hybridization stringency was achieved by washing the membranes in hybridization wash (150 mM NaCl, 20 mM Tris (pH 7.4), 6 mM EDTA and 0.01% SDS) for two 10-min intervals at experimentally determined temperatures of dissociation (T_d) specific for the individual sets of probes (SAR11, 55 °C; Bacteria, 50 °C; 338F, 50 °C). Nucleic acid staining was achieved by transferring the membrane to hybridization wash containing 5 µg ml⁻¹ DAPI for 10 min at 4 °C. DAPI was rinsed off for 2 min in a final hybridization wash at 4 °C. All reagents were filtered through 0.2-µm membranes.

Fluorescent microscopy

After mounting the filters in Citifluor (Ted Pella), Cy3-positive and DAPI-positive cells were counted for each field of view using a Leica DMRB epifluorescence microscope equipped with a Hamamatsu ORCA-ER CCD digital camera, filter sets appropriate for Cy3 and DAPI, and Scanalytics IPLab v3.5.5 scientific imaging software. Consistent exposure times of 1 and 5 s were used for DAPI and Cy3 images, respectively. We segmented DAPI images using IPLab software and overlaid them on corresponding Cy3 image segmentations to identify positive probe signals with corresponding DAPI signals.

Negative control counts were determined using the same technique and subtracted from positive probe counts to correct for autofluorescence and nonspecific binding. We counted an average of 693 DAPI-staining cells for surface samples from 1 to 150 m ($n = 72$), and an average of 320 DAPI staining cells for mesopelagic samples from 250 to 3,000 m ($n = 21$).

Bulk nucleic acid hybridization

Unamplified small subunit ribosomal RNA was probed essentially as described²⁷. We used only two oligonucleotide probes, SAR11-441R and SAR11-542R, from the *in situ* set of probes because these two probes could detect nearly all representative SAR11 clade 16S rRNA sequences and because the additive effect on intensity needed to identify small cells using FISH was not necessary. Stringency conditions used for this study were determined empirically. Individually determined hybridization wash temperatures were 42 °C for the SAR11 probes, making it possible to combine the two SAR11 probes used in this study into a single probe set. The SAR11 signal was compared with the total bacterial RNA signal obtained using probe EUB-338Rpl and the specific hybridization value was determined by comparison with the SAR11 positive control strains, HTCC1040 and HTCC1062 (ref. 22).

Received 31 May; accepted 7 October 2002; doi:10.1038/nature01240.

- Giovannoni, S. & Rappé, M. *Microbial Ecology of the Oceans* (ed. Kirchman, D. L.) 47–84 (John Wiley and Sons, New York, 2000).
- Cho, B. C. & Azam, F. Major role of bacteria in biogeochemical fluxes in the ocean's interior. *Nature* **332**, 441–443 (1988).
- Fuhrman, J. A., Sleeter, T. D., Carlson, C. A. & Proctor, L. Dominance of bacterial biomass in the Sargasso Sea and its ecological implications. *Mar. Ecol. Prog. Ser.* **57**, 207–217 (1989).
- Buck, K. R., Chavez, F. P. & Campbell, L. Basin-wide distributions of living carbon components and the inverted trophic pyramid of the central gyre of the North Atlantic Ocean, summer 1993. *Aquat. Microb. Ecol.* **10**, 238–298 (1996).
- Carlson, C. A., Ducklow, H. W. & Sleeter, T. D. Stocks and dynamics of bacterioplankton in the northwestern Sargasso Sea. *Deep-Sea Res.* **II** **43**, 491–516 (1996).
- Karner, M. B., Delong, E. F. & Karl, D. M. Archaeal dominance in the mesopelagic zone of the Pacific Ocean. *Nature* **409**, 507–510 (2001).
- Fuhrman, J. A. & Ouverney, C. C. Marine microbial diversity studied via 16S rRNA sequences: cloning results from coastal waters and counting native archaea with fluorescent single cell probes. *Aquat. Ecol.* **32**, 3–15 (1998).
- Wintzingerode, F. V., Gobel, U. B. & Stackebrandt, E. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol. Rev.* **21**, 213–229 (1997).
- Hicks, R. E., Amann, R. I. & Stahl, D. A. Dual staining of natural bacterioplankton with 4′-6-diamidino-2-phenylindole and fluorescent oligonucleotide probes targeting kingdom-level 16S rRNA sequences. *Appl. Environ. Microbiol.* **58**, 2158–2163 (1992).
- DeLong, E. F., Wickham, G. S. & Pace, N. R. Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science* **243**, 1360–1363 (1989).
- Amann, R. I., Krumholz, L. & Stahl, D. A. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *J. Bacteriol.* **172**, 762–770 (1990).
- Glöckner, F. O. *et al.* An *in situ* hybridization protocol for detection and identification of planktonic bacteria. *Syst. Appl. Microbiol.* **19**, 403–406 (1996).
- Frischer, M. E., Floriani, P. J. & Nierzwicki-Bauer, S. A. Differential sensitivity of 16S rRNA targeted oligonucleotide probes used for fluorescence *in situ* hybridization is a result of ribosomal higher order structure. *Can. J. Microbiol.* **42**, 1061–1071 (1996).
- Fuchs, B. M. *et al.* Flow cytometric analysis of the *in situ* accessibility of *Escherichia coli* 16S rRNA for fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.* **64**, 4973–4982 (1998).
- Fuchs, B. M., Glöckner, F. O., Wulf, J. & Amann, R. Unlabeled helper oligonucleotides increase the *in situ* accessibility to 16S rRNA of fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.* **66**, 3603–3607 (2000).
- Lee, S., Malone, C. & Kemp, P. F. Use of multiple 16S rRNA-targeted fluorescent probes to increase signal strength and measure cellular RNA from natural planktonic bacteria. *Mar. Ecol. Prog. Ser.* **101**, 193–201 (1993).
- Schönhuber, W., Fuchs, B., Juretschko, S. & Amann, R. I. Improved sensitivity of whole-cell hybridization by the combination of horseradish peroxidase-labeled oligonucleotides and tyramide signal amplification. *Appl. Environ. Microbiol.* **63**, 3268–3273 (1997).
- Pernthaler, A., Preston, C. M., Pernthaler, J., DeLong, E. F. & Amann, R. Comparison of fluorescently labeled oligonucleotide and polynucleotide probes for the detection of pelagic marine bacteria and archaea. *Appl. Environ. Microbiol.* **68**, 661–667 (2002).
- Sieracki, M. E., Johnson, P. W. & Sieburth, J. M. Detection, enumeration, and sizing of planktonic bacteria by image-analyzed epifluorescence microscopy. *Appl. Environ. Microbiol.* **49**, 799–810 (1985).
- Porter, K. G. & Feig, Y. S. The use of DAPI for identifying and counting aquatic microflora. *Limnol. Oceanogr.* **25**, 943–948 (1980).
- Kemp, P. F., Lee, S. & LaRoche, J. Estimating the growth rate of slowly growing marine bacteria from RNA content. *Appl. Environ. Microbiol.* **59**, 2594–2601 (1993).
- Rappé, M. S., Connon, S. A., Vergin, K. L. & Giovannoni, S. J. Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418**, 630–633 (2002).
- Cottrell, M. T. & Kirchman, D. L. Community composition of marine bacterioplankton determined by 16S rRNA gene clone libraries and fluorescence *in situ* hybridization. *Appl. Environ. Microbiol.* **66**, 5116–5122 (2000).
- Whitman, W. B., Coleman, D. C. & Wiebe, W. J. Prokaryotes: The unseen majority. *Proc. Natl Acad. Sci. USA* **95**, 6578–6583 (1998).
- Field, K. G. *et al.* Diversity and depth-specific distribution of SAR11 cluster rRNA genes from marine planktonic bacteria. *Appl. Environ. Microbiol.* **63**, 63–70 (1997).
- Ludwig, W. *et al.* Bacterial phylogeny based on comparative sequence analysis. *Electrophoresis* **19**, 544–568 (1998).
- Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63 (1990).

Acknowledgements We thank R. Parsons, N. Nelson, the BATS scientific team and officers and crew of the RV *Weatherbird II* for help with collecting and processing samples. This work was supported by grants from Oregon State University, the Murdock Charitable Trust and the National Science Foundation.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to S.J.G. (e-mail: steve.giovannoni@orst.edu).

Macroevolution simulated with autonomously replicating computer programs

Gabriel Yedid*† & Graham Bell*

* Biology Department, McGill University, 1205 avenue Dr Penfield, Montreal, Quebec, Canada H3A 1B1

† The Center for Microbial Ecology, 540 Plant and Soil Sciences Building, Michigan State University, East Lansing, Michigan 48824-1325, USA

The process of adaptation occurs on two timescales. In the short term, natural selection merely sorts the variation already present in a population, whereas in the longer term genotypes quite different from any that were initially present evolve through the cumulation of new mutations. The first process is described by the mathematical theory of population genetics. However, this theory begins by defining a fixed set of genotypes and cannot provide a satisfactory analysis of the second process because it does not permit any genuinely new type to arise. The evolutionary outcome of selection acting on novel variation arising over long periods is therefore difficult to predict. The classical problem of this kind is whether 'replaying the tape of life' would invariably lead to the familiar organisms of the modern biota^{1,2}. Here we study the long-term behaviour of populations of autonomously replicating computer programs and find that the same type, introduced into the same simple environment, evolves on any given occasion along a unique trajectory towards one of many well-adapted end points.

When replicate lines of microbes are cultured in the same conditions, they usually evolve higher rates of growth in parallel. In principle, this need involve no more than the successive substitution of the same series of mutations and thus the emergence of the same multilocus genotype^{2,3}. The tempo of adaptation can be analysed through the theory of periodic selection^{4,5}; the lack of substantial genetic variance of fitness among replicate lines suggests that this simple description might often be adequate^{6,7}. In a few cases where the relevant loci are known this has been confirmed directly⁸⁻¹⁰, but in other experiments replicate lines have diverged in fitness¹¹, or the pattern of correlated responses has shown that they have diverged genetically, despite their similar direct responses^{12,13}. In some cases, the physiological and genetic responses occurring in replicate lines have been identified, showing how the same selection pressure can lead to quantitatively and qualitatively different outcomes^{14,15}. In experiments with large multicellular organisms such as *Drosophila*, differences among replicate lines can be caused by initial differences among the founding populations, or by inbreeding or drift during the course of the experiment^{16,17}. In microbial experiments with large populations of asexual organisms derived from isogenic founders, these explanations are ruled out, raising the possibility that long-term evolutionary change is often highly contingent.

We studied an auto-adaptive genetic system called Tierra^{18,19}, comprising a virtual microcosm inhabited by self-replicating computer programs ('creatures'). The genome of each creature is a sequence of instructions that directs the flow of control. Each instruction performs some elementary operation, such as incrementing a register or moving the instruction pointer to another location in the genome. When an appropriate sequence of instructions is correctly executed, the sequence is copied elsewhere in computer memory; changing the sequence of instructions usually changes the rate of replication. Selection between the variants produced by mutation is caused by competition for CPU time: variants that replicate more rapidly tend to increase in frequency. The evolving population of programs thus resembles an evolving population of bacteria or viruses, and such auto-adaptive genetic systems are now coming into use to supplement conventional analytical methods in evolutionary theory²⁰. We have shown previously that the microevolutionary dynamics of Tierran populations are adequately described by conventional population genetics theory²¹.

We began each experiment with a single creature that was allowed to proliferate indefinitely, with an initial carrying capacity of 500 individuals and a mutation rate of 0.1 per genome per replication.

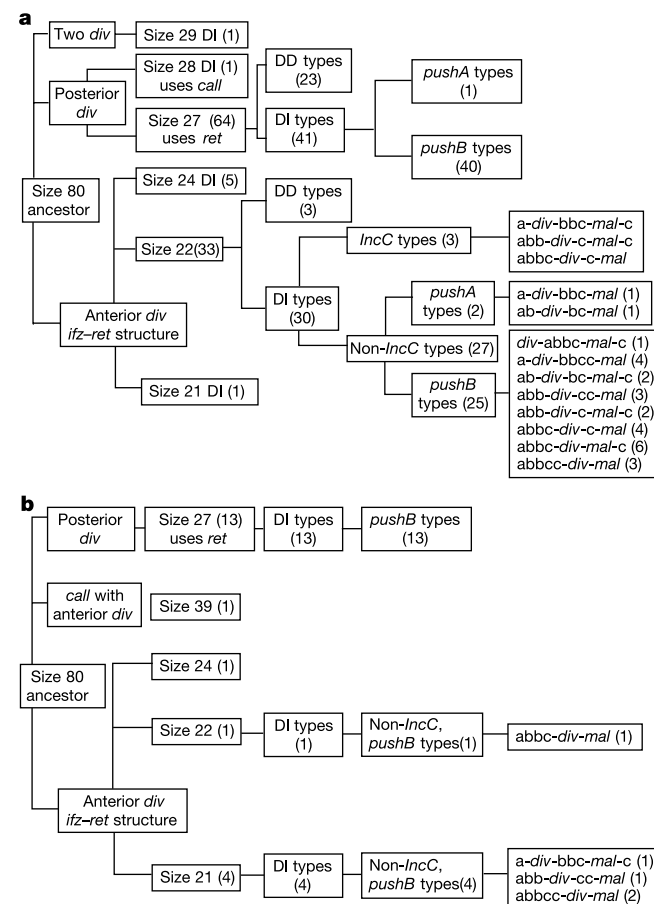


Figure 1 Variety of evolutionary end points. Genotypes are classified according to the presence and position of key instructions such as *div* (divide), *mal* (allocate memory), *IncC* (increment register C), *PushX* (put a value in register X onto a stack), *ifz* (if register equals zero then execute next instruction, else advance two instructions) and *ret* (return). DD types are density-dependent and cannot replicate as single copies in pure culture; DI types are density-independent. The diversity of size-22 DI types is arranged according to the position of instructions relative to *div* and *mal*. **a**, is *adrb* or *adra*; **b**, is *subAAC* or *movBA*; **c**, is *subCAB* or *pushB*. The numbers in parentheses give the number of observed outcomes of a given kind. **a**, Small-population experiments; **b**, large-population experiments.

The environment was the simplest possible, in which individuals competed for CPU time alone and other biotic interactions were not permitted. The process was continued until no further genetic change could be discerned. In each case the population at the end point was fixed for a single genotype (with a certain amount of neutral variation). We collected these genotypes for 109 independent experiments.

Selection for more rapid replication was effective in all cases, causing a reduction in the number of instructions executed by a creature in the course of replication. As expected, this was accompanied by a correlated decrease in genome size. The ancestral genome of 80 instructions was eventually replaced by much shorter sequences of fewer than 30 instructions, reflecting a decrease in the number of instructions executed per replication (I_E) from 827 I_E in the ancestor to as few as 135 I_E in its remote descendants. The direct phenotypic response to selection was thus mostly, although not completely, in parallel (the mean among small-population lines was 177.4 I_E , s.d. 25.8 I_E).

In contrast, the genotypic basis of the response was highly divergent, with several structurally different genomes emerging as possible end points of evolution (Fig. 1a). The two main groups of genotypes comprised either 27 or 22 instructions and were qualitatively different in structure. Genomes of size 27 invariably had the *div* (divide by copying the parental sequence into the memory assigned) instruction as close as possible to the end of the sequence, whereas in size-22 genomes *div* was in the front half of the sequence, often very near the beginning. The size-27 genomes were rather uniform, with most of them belonging to a single type, whereas the size-22 genomes were highly diverse. They often differed with regard to the presence or position of important instructions that direct the flow of control; they also differed with regard to the genetic organization of the region around the key instructions *div* and *mal* (allocate a block of memory into which the offspring sequence can be copied). Furthermore, both size-27 and size-22 genomes had density-dependent and density-independent versions. The density-independent versions are normal self-replicating algorithms, but the density-dependent versions are unable to replicate as single copies and require the presence of a similar sequence nearby. Finally, several other end points were also observed, including bizarre genotypes such as one bearing two copies of *ret* (return from the copy sequence and reinitiate the sequence of events leading to replication). The microscopic details of these genotypes are in themselves of little interest, but the manner in which epistatic and social interactions affected evolutionary outcomes suggests the existence of general principles that might apply to all systems of self-replicators.

The time taken to reach the end point also differed between replicate experiments. Even for the rather uniform size-27 density-dependent sequence was the end point.

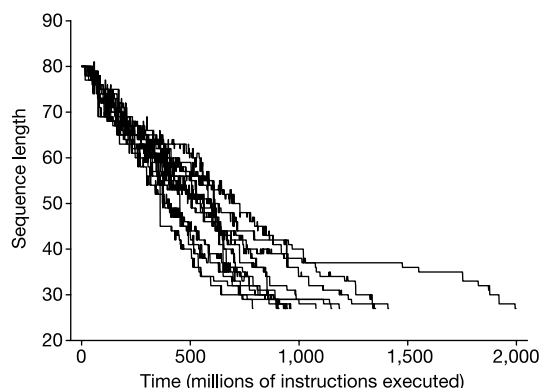


Figure 2 Change in sequence length through time of experiments for which a size-27 density-dependent sequence was the end point.

one aberrant case in which the population remained almost static for 1,000 million instructions (Fig. 2). Moreover, the nature of the intermediate genotypes was different in every experiment. This wide range of trajectories indicates that the genetic background in which mutations arise affects the route to the end point and the time taken to reach it, an observation consistent with the assertion that mutational order is important in causing divergence between populations²².

The trajectory itself was studied in more detail through 'go-back' experiments, in which an intermediate genotype was recovered and used to initiate a new evolutionary process. At the beginning of any experiment the founding genotype is 'totipotent' and can evolve towards any end point. As time goes by, the sequence of changes that its descendants have accumulated tends to restrict the range of end points that it can attain, until eventually, near the end of the process, it is nearly certain to reach a particular end point. In the simple Tierra system, the crucial fate-determining step is the placement of *div* either at the rear or near the front of the genome, which usually directs subsequent evolution towards either size-22 or size-27 genotypes. This step depends on the epistatic interaction between an *ifz-ret* pair situated in the copy procedure, near the middle of the genome, with *div*. (The *ifz* instruction can cause the instruction pointer to move forward two places, rather than a single place.) The *ifz-ret* combination is normally strongly deleterious, but increases replication rate in conjunction with an anterior *div* (see Methods). When this combination arises, whether early or late in the history of a lineage, evolution usually leads to one of the short sequences that retain an anterior *div*; otherwise, it tends towards one of the longer sequences with a posterior *div* (see Methods).

Long-term evolutionary dynamics depend on the flux of beneficial mutations, which is determined by population size N and mutation rate per site u (refs 23, 24). If $Nu > 1$ then most mutations, beneficial or not, occur at once, the most fit are substituted simultaneously, and replicate populations evolve in parallel; if $Nu \ll 1$ then beneficial mutations are substituted successively and replicate populations diverge because each experiences a unique sequence of beneficial mutations. In our experiments the genome size is on average about 50, so that $Nu \approx 500 \times 0.1/50 \approx 1$. Moreover, beneficial mutations are less frequent: the frequency of mutations of small effect that increase fitness is about $0.35u$, and that of mutations of large effect less than $0.02u$ (ref. 21). The extensive divergence that we observed might have been attributable to a restricted supply of mutations caused by small population size. We therefore conducted a further series of 20 experiments with $N = 10,000$ size-80 creatures and thus $Nu \approx 20$. The predominance of size-27 end points was more marked and they were, as before, rather uniform in structure. Nevertheless, a number of smaller types of size 21–24 evolved, together with a highly aberrant size-39 creature that incorporated a sequence (involving *call* in conjunction with an anterior *div* and lacking the conventional 'head' sequence seen in all other types) never before seen in our experiments (Fig. 1b). Thus, evolution continues to be highly contingent even at relatively high values of Nu , because of the strongly epistatic nature of fitness. In an infinite population, evolution would in principle be completely predictable and populations would not diverge permanently, because in an infinite population not only would any combination of mutations occur but each would be infinitely abundant. However, in real populations only a very small fraction of possible combinations of mutations occur in any generation, and moderate degrees of epistasis ensure that replicate populations often diverge towards different end points.

Our experiments indicate that contingency might be a general feature of evolution in systems with unbounded, nearly neutral variation. It is possible to construct complex environments in which, for example, programs are rewarded for their ability to manipulate integers^{25,26}. In such conditions it might not be very

surprising to learn that many effective solutions exist and that populations consequently diverge in both phenotype and genotype. However, the situation represented in these experiments is the simplest possible: directional selection acting on beneficial mutations that reduce genome size, leading to strong phenotypic convergence and the eventual fixation of a single genotype. The situation is strikingly similar to the evolution of the virus Q β *in vitro*, where it evolves strongly reduced genome size and increased replication rate²⁷. Nevertheless, we have demonstrated not only that the genetic basis of evolutionary change varies between replicate realizations but also that the evolutionary fate of evolving lineages becomes steadily more closely determined through time and is strongly influenced by key innovations involving irreversible epistatic interactions between loci. Thus, even in this simplest of situations, the outcome of evolution can be any of several genotypes, and it cannot be predicted which will emerge in any particular realization. □

Methods

All experiments were conducted using Tierra v.5.0, and run on computers based on Intel Pentium II and Pentium III running the Microsoft Windows 98 and Windows 2000 operating systems. The end points of the small-population and large-population experiments can be downloaded from McGill University's website (<http://www.mcgill.ca/biology/faculty/bell>).

Basic experiments and 'go-back' experiments

A single copy of the Tierran founder genotype, 80aaa, was used to seed a series of 109 small-population and 20 large-population replicates conducted with the methodology of Yedid and Bell²¹. Evolution was considered to have ceased when the only replication-competent genotypes being produced were nearly neutral variants and no further change in size or fitness seemed possible. Genotypes from two experiments that resulted in the two major products of evolution (27 density-independent (DI) and 22 DI types) were selected for a series of 'go-back' experiments. We extracted a dominant genotype from points near the end, middle and beginning of each experiment, and inoculated the system with that genotype alone, using the same culture conditions as for the original experiment. Ten replicates were run for each genotype. This protocol was repeated for the 20 large-population experiments, except that the starting carrying capacity was 10,000 size-80 creatures.

Test of fitness effects of *div/ifz-ret* combination

Two genotypes from an experiment that resulted in a size 22 DI end point, one of size 56 lacking the *div/ifz-ret* fingerprint and a second of size 54 containing the fingerprint, were selected to investigate whether or not occurrence of the fingerprint by itself conferred enhanced fitness. Dummy instructions were added to a non-critical section of the genome of the size-54 genotype to slow its rate of replication, resulting in a size-57 genotype equal in fitness to the size-56 type. The two genotypes were competed against one another, saving all progeny genotypes so as to trace unequivocally the lineage of subsequent dominant types; the added instructions provided a marker for identifying progeny of the size-54/57 genotype. One set of experiments was inoculated with a single individual of each type at low density; a second set was inoculated with high and equal numbers of each. In the great majority (>90%) of the tests conducted, dominant genotypes derived from the size-54/57 type ultimately prevailed, even in cases where progeny of the size-56 type initially had a short-term advantage.

Determination of end point by *div/ifz-ret* combination

A size-63 genotype from an experiment that resulted in a size-27 DI end point was modified to contain the *div/ifz-ret* fingerprint, and used to inoculate the system as in the 'go-back' experiments. It was found in all attempts that the resulting end points were smaller than size 27 (mostly of size 22) and retained the *div/ifz-ret* fingerprint. A second set of experiments was conducted in which the same size-63 genotype was modified with only the anterior *div* near the creature's head, to examine whether the presence of this instruction was a necessary preadaptation for genotypes smaller than size 27 to occur. Wherever it was retained, a genotype smaller than size 27 was the eventual end point, although the anterior *div* was lost through mutation roughly half the time, resulting in size-27 types with the posterior *div-ret* combination.

Received 14 June; accepted 9 September 2002; doi:10.1038/nature01151.

- Gould, S. J. *Wonderful Life: The Burgess Shale and the Nature of History* (W. W. Norton, New York, 1989).
- Dykhuizen, D. Experimental studies of natural selection in bacteria. *Annu. Rev. Ecol. Syst.* **21**, 373–398 (1990).
- Fisher, R. A. *The Genetical Theory of Natural Selection* (Dover, New York, 1930).
- Atwood, K. C., Schneider, L. K. & Ryan, F. J. Selective mechanisms in bacteria. *Cold Spring Harb. Symp. Quant. Biol.* **16**, 345–355 (1951).
- Gerrish, P. J. The rhythm of microbial adaptation. *Nature* **413**, 299–302 (2001).
- Dykhuizen, D. & Hartl, D. L. Evolution of competitive ability in *Escherichia coli*. *Evolution* **35**, 581–594 (1981).
- Travisano, M., Mongold, J. A., Bennett, A. F. & Lenski, R. E. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science* **267**, 87–90 (1995).

- Kramer, F. R., Mills, D. R., Cole, P. E., Nishihara, T. & Spiegelman, S. Evolution *in vitro*: sequence and phenotype of a mutant RNA resistant to ethidium bromide. *J. Mol. Biol.* **89**, 719–736 (1974).
- Bull, J. J. *et al.* Exceptional convergent evolution in a virus. *Genetics* **147**, 1497–1507 (1997).
- Nakatsu, C. H. *et al.* Parallel and divergent genotypic evolution in experimental populations of *Ralstonia* sp. *J. Bacteriol.* **180**, 4325–4331 (1998).
- Lenski, R. & Travisano, M. Dynamics of adaptation and diversification: a 10,000 generation experiment with bacterial populations. *Proc. Natl Acad. Sci. USA* **91**, 6808–6814 (1994).
- Korona, R., Nakatsu, C. H., Forney, L. J. & Lenski, R. E. Evidence for multiple adaptive peaks from populations of bacteria evolving in a structured habitat. *Proc. Natl Acad. Sci. USA* **91**, 9037–9041 (1994).
- Travisano, M. Long-term experimental evolution in *Escherichia coli*. VI. Environmental constraints on adaptation and divergence. *Genetics* **146**, 471–479 (1997).
- Hartley, B. S. Experimental evolution of ribitol dehydrogenase. *Microorganisms as Model Systems for Studying Evolution* (ed. Mortlock, R. P.) 23–54 (Plenum, New York, 1984).
- Yin, J. Evolution of bacteriophage T7 in a growing plaque. *J. Bacteriol.* **175**, 1272–1277 (1993).
- Cohan, F. M. & Hoffmann, A. A. Genetic divergence under uniform selection. II. Different responses to selection for knockdown resistance to ethanol among *Drosophila melanogaster* populations and their replicate lines. *Genetics* **114**, 145–163 (1986).
- Hoffmann, A. A. & Harshman, L. G. Desiccation and starvation resistance in *Drosophila*: Patterns of variation at the species, population and intrapopulation levels. *Heredity* **83**, 637–643 (1999).
- Ray, T. S. An approach to the synthesis of life. *Artificial Life II, Santa Fe Institute Studies in the Sciences of Complexity* (eds Farmer, D. J., Langton, C., Rasmussen, S. & Taylor, C.) **Vol. 11** 371–408 (Addison Wesley, Redwood City, California, 1991).
- Ray, T. S. Tierra v. 5.0, with documentation. Available via anonymous ftp from the Tierra home page at <http://www.isd.atr.co.jp/~ray/tierra>
- Wilke, C. O., Wang, J. L., Ofria, C., Lenski, R. E. & Adami, C. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature* **412**, 331–333 (2001).
- Yedid, G. & Bell, G. Microevolution in an electronic microcosm. *Am. Nat.* **157**, 465–487 (2001).
- Mani, G. S. & Clarke, B. C. Mutational order: a major stochastic process in evolution. *Proc. R. Soc. Lond. B* **240**, 29–37 (1990).
- Johnson, P. A., Lenski, R. E. & Hoppensteadt, F. C. Theoretical analysis of divergence in mean fitness between genetically identical populations. *Proc. R. Soc. Lond. B* **259**, 125–130 (1995).
- Wahl, L. M. & Krakauer, D. C. Models of experimental evolution: the role of genetic chance and selective necessity. *Genetics* **156**, 1437–1448 (2000).
- Adami, C. Learning and complexity in genetic auto-adaptive systems. *Physica D* **80**, 154–170 (1995).
- Adami, C., Ofria, C. & Collier, T. C. Evolution of biological complexity. *Proc. Natl Acad. Sci. USA* **97**, 4463–4468 (2000).
- Spiegelman, S. An approach to the experimental analysis of precellular evolution. *Q. Rev. Biophys.* **4**, 213–253 (1971).

Acknowledgements This research was supported by a Research Grant from the Natural Sciences and Engineering Research Council of Canada to G.B.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.B. (e-mail: graham.bell@mcgill.ca).

Long-term dendritic spine stability in the adult cortex

Jaime Grutzendler, Narayanan Kasthuri & Wen-Biao Gan

Molecular Neurobiology Program, Skirball Institute, Department of Physiology and Neuroscience, New York University School of Medicine, 540 First Avenue, New York, New York 10016, USA

The structural dynamics of synapses probably has a crucial role in the development and plasticity of the nervous system. In the mammalian brain, the vast majority of excitatory axo-dendritic synapses occur on dendritic specializations called 'spines'. However, little is known about their long-term changes in the intact developing or adult animal. To address this question we developed a transcranial two-photon imaging technique to follow identified spines of layer-5 pyramidal neurons in the primary visual cortex of living transgenic mice expressing yellow fluorescent protein. Here we show that filopodia-like dendritic protrusions, extending and retracting over hours, are abundant in young animals but virtually absent from the adult. In young mice, within the 'critical period' for visual cortex development, ~73% of spines remain stable over a one-month interval; most changes are associated with spine elimination. In contrast, in adult mice, the overwhelming majority of spines (~96%) remain

stable over the same interval with a half-life greater than 13 months. These results indicate that spines, initially plastic during development, become remarkably stable in the adult, providing a potential structural basis for long-term information storage.

Changes in synaptic connections occur in various neurological processes ranging from the development of neuronal circuitry, learning and memory, to injury-related recovery^{1–6}. In the mammalian nervous system, synaptic connections are rapidly formed and eliminated in early postnatal life, resulting in a functionally wired mature brain^{1–3,7,8}. Classic studies have demonstrated time windows of heightened functional and anatomical plasticity, called ‘critical periods’, during which the developing nervous system can be rapidly and permanently modified by experience^{1,9}. In the visual cortex, for example, brief monocular deprivation during a critical period in development, but not in the adult animal, causes lifelong changes in synaptic connections^{1,9,10}. The structural basis underlying this transition from a plastic to a more stable adult nervous system is poorly understood. In adulthood, although the nervous system retains a certain capacity to reorganize after peripheral and central alterations of inputs^{4–6,11–13}, the degree of structural plasticity in the intact adult brain remains largely unknown.

To study the dynamic nature of synaptic structures in the intact central nervous system, we developed a technique that allows long-term imaging of individual dendritic spines in the living mouse by two-photon microscopy (Fig. 1a, b; see also Methods). We used transgenic mice expressing yellow fluorescent protein (YFP) in pyramidal neurons, predominantly in layer 5 of the cortex¹⁴. As shown in Fig. 1, individual spines and filopodia on branches of apical dendrites in the visual cortex were imaged at high optical resolution through a thin-skull window. The presence of a thinned skull did not seem to interfere with our ability to resolve the thinnest dendritic protrusions such as filopodia (see images with and without skull, as well as high-resolution confocal images of fixed brain slices in Supplementary Information). Dendrites were then re-located at subsequent time points by using as landmarks surrounding meningeal blood vessels and low-magnification image stacks of nearby neuronal processes (Fig. 1a, b).

We began by imaging segments of apical dendrites, typically in layers I and II, of the primary visual cortex over periods of hours in 1-month-old mice. In these young animals we observed two distinct populations of dendritic protrusions: filopodium-like structures (long and thin processes without spine heads) and spines (Figs 1c, d and 2; see also Supplementary Information). On the basis of their length and the ratios between length, head and neck diameter (see Methods), we classified protrusions as either filopodia or spines and found that the ratio of filopodia to spines was $\sim 1:8.4$ at this age ($n = 1,580$ protrusions, 10 animals; Fig. 1e). Although filopodia, like dendritic spines, can form synaptic contacts^{15,16}, previous studies suggest that they are more dynamic than spines and are actively involved in synaptogenesis^{17–20}. Consistent with previous studies^{17–20} was our observation that the vast majority of filopodia ($90.5 \pm 8.1\%$, $n = 60$, two animals) extended or retracted rapidly, with some eliminated and others formed *de novo* during the 4-h observation period (Fig. 1c, d and f). In contrast, only 5 of 476 spines ($\sim 1.0\%$) extended or retracted during the same period (Fig. 1f).

To further examine long-term changes in filopodia and spines in young (1-month-old) animals, identified dendritic segments were imaged over intervals ranging from 3 days to 1 month. We found that most pre-existing filopodia ($86.2 \pm 12.0\%$) were eliminated, and many new ones ($109.5 \pm 36.7\%$) were formed within 3 days, indicating a rapid turn-over of filopodia at this age (Fig. 2e). Because the ratio of filopodia to spines decreased significantly within the next 2–4 weeks (Fig. 1e), the filopodia that formed within this period must have been either converted to spines or eliminated. Indeed, time-lapse imaging over the next 2–4 weeks showed that the vast majority of filopodia were eliminated, and only rarely did they seem to convert into spines ($<1\%$; Fig. 2a, b and e). Because the rate of new spine formation is also very low at this age (see below), we conclude that the vast majority of filopodia between the first and second months of postnatal life do not contribute to the creation of new connections.

In contrast to filopodia, spines were much more stable at this young age, with $94.2 \pm 1.5\%$ ($n = 365$, three animals) remaining

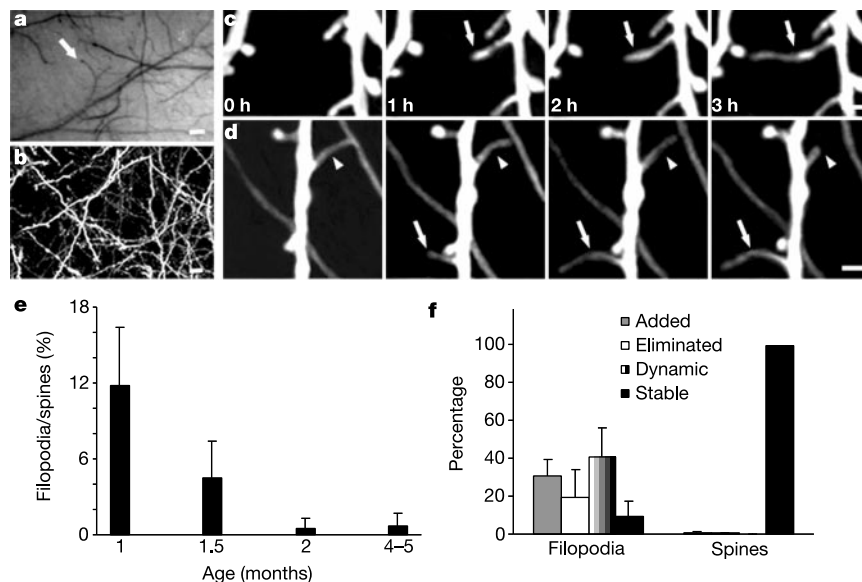


Figure 1 Transcranial two-photon imaging in young animals shows dynamic dendritic filopodia. **a**, Charge-coupled device camera view of the vasculature below the thinned skull. The arrow indicates the region where subsequent two-photon images were obtained. **b**, Two-dimensional projection of a three-dimensional stack of dendritic branches and axons in the primary visual cortex. **c**, **d**, Time-lapse imaging (one hour intervals) in two different 1-month-old animals revealed that filopodia underwent rapid extension (arrows) and retraction (arrowheads), whereas spines on the same dendritic

branches remained stable. **e**, Percentage of filopodia (number of filopodia as a percentage of number of spines) decreases as a function of age. **f**, Percentage of filopodia and spines added (number of filopodia or spines added as a percentage of the pre-existing number of filopodia or spines), eliminated and showing either dynamic changes or stability over a period of 4 h. Data are presented as means $\pm$ s.d. Scale bars, 50 μm (**a**), 5 μm (**b**) and 1 μm (**c** and **d**).

stable over 3 days. Moreover, $81.6 \pm 2.7\%$ ($n = 359$, three animals) remained stable over 2 weeks, and $73.1 \pm 2.2\%$ ($n = 361$, three animals) remained stable over a 1-month interval (Fig. 2c, d and f). Interestingly, the percentage of spine elimination ($26.9 \pm 2.2\%$) was significantly higher than that of spine formation ($3.0 \pm 2.9\%$, $P < 0.0001$; t -test) over 1 month, indicating that the net number of spines decreases during this period. This is consistent with previous studies in fixed preparations demonstrating a significant net loss of synapses during the development of monkey and human visual cortex^{7,21,22}.

Studies on fixed preparations also indicate that synaptic density and the total number of spines remain relatively stable in adulthood^{17,8,21–23}, suggesting that adult spines are either very stable or are undergoing constant turnover with comparable rates of spine formation and elimination. To determine the degree of spine plasticity in adulthood, we followed dendrites in animals of average age 4.2 ± 0.8 months (14 animals) over intervals ranging from hours to months. Consistent with studies both *in vitro* and in fixed preparations showing a developmental downregulation of rapidly motile filopodia^{17–20}, we found an almost complete absence of filopodia in adult mice (5 of 630 protrusions, five animals; Fig. 1e and see also Supplementary Information). Furthermore, time-lapse imaging showed that only 2 of 208 dendritic protrusions (two animals) were identified anatomically as filopodia and underwent morphological changes during a 4-h observation period. Neither extension nor retraction was found in 206 spines examined over the same interval. Because filopodia are scarce in adults ($<1\%$), we did not classify dendritic protrusions into filopodia and spines in our subsequent studies of long-term changes in adults.

When imaged over a 3-day interval, we found that in adults, the number and location of spines remained essentially unchanged ($99.3 \pm 0.7\%$, $n = 408$, three animals; Fig. 3a–d and m). Remarkably, when the viewing interval was extended to 1 month, $96.5 \pm 2.2\%$ of spines were still stably maintained (3.5% eliminated) and the number of new spines was only $3.2 \pm 2.1\%$ of the pre-existing ones ($n = 389$, five animals; Fig. 3m). Over a 2-month period, $92.4 \pm 1.4\%$ of spines remained stable, and $5.5 \pm 2.0\%$ were added

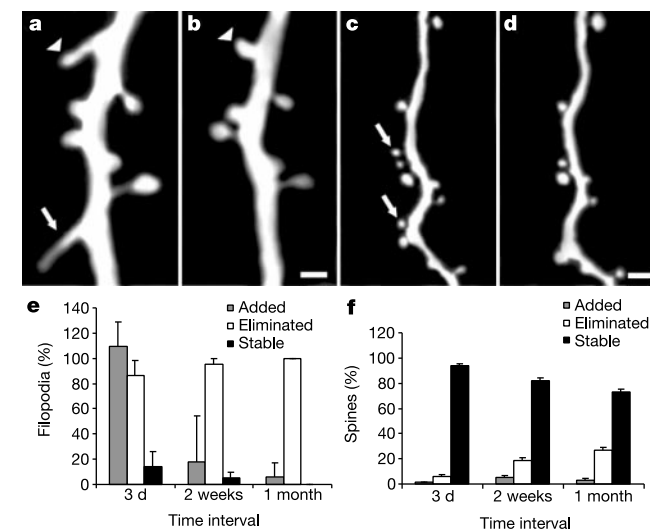


Figure 2 Dendrites in young (1-month-old) animals show predominantly spine and filopodia elimination over long intervals. **a, b**, Repeated imaging of dendritic branches over a 1-month interval revealed one filopodium that retracted (arrow) and another that converted into a new spine (arrowheads) by the second view (**b**). **c, d**, Repeated imaging of dendritic branches taken 1 month apart revealed the elimination of two spines (arrows) by the second view (**d**). **e**, Percentage of filopodia added (number of filopodia added as a percentage of the pre-existing number of filopodia), eliminated or stable as a function of viewing interval. **f**, Percentage of spines added, eliminated or stable as a function of viewing interval. Data are presented as means $\pm$ s.d. Scale bar, 1 μ m.

($n = 437$, four animals; Fig. 3e–h and m). In two animals imaged over a 4-month period, $80.4 \pm 2.0\%$ of spines remained stable and $9.5 \pm 3.8\%$ were added ($n = 101$). Finally, even older animals (10 months old) showed similar rates of spine elimination and formation over a 1-month interval (Fig. 4a). Thus, spines are remarkably stable over long periods in the adult.

Although the number and distribution of adult spines remained largely constant, many spines did undergo marked changes in length or head diameter over 3-day and 1-month periods (Fig. 3i–l). However, we did not observe a trend towards spine enlargement or reduction over either interval ($n = 50$ over 3 days, and $n = 48$ over 1 month; $P > 0.3$; t -test). The occurrence of change in the length or diameter of spines was greater during a 1-month interval than over a 3-day interval (F -test of variance, $P < 0.000024$ for spine length and $P < 0.0037$ for spine diameter), suggesting that these morphological changes accumulate progressively with time in adult mice and are unlikely to be due to rotational or movement artefacts. Because spine diameter and length are well correlated with the size of the postsynaptic density and the magnitude of signals transmitted to the dendritic shaft^{24,25}, changes in spine morphology could therefore provide a mechanism of modulating synaptic efficacy.

As shown in Fig. 4a, the rate of spine elimination decreases during post-natal development, whereas the rate of spine formation remains largely unchanged across different ages. On the basis of the observed rates of spine elimination and formation in animals

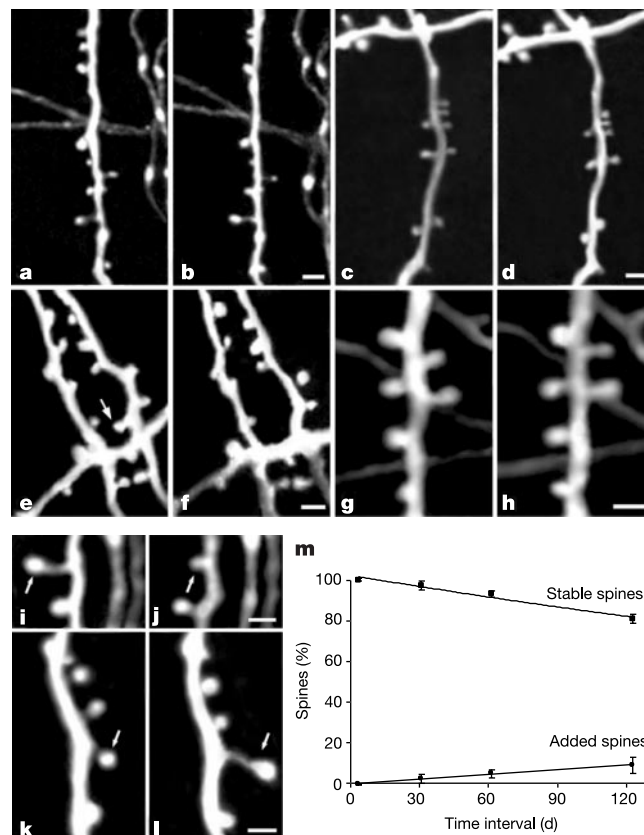


Figure 3 Dendritic spines in adult mice demonstrate long-term stability. **a** and **b**, **c** and **d**, Dendritic branches from two animals imaged 3 d apart show the same spines at the same locations. **e** and **f**, **g** and **h**, Repeated imaging of dendritic branches from two animals (4 months old) imaged 2 months apart show little change in spine number or location. The arrow in **e** points to a spine that is eliminated in the second view (**f**). **i** and **j**, **k** and **l**, Repeated imaging in adult animals shows spine morphology changes over a 1-month interval (see Methods). Spine length and head diameter were reduced markedly (**i**, **j**) or substantially increased by the second view (**k**, **l**). **m**, Percentage of spines that remained stable or were added as a function of imaging interval for adult animals (4.2 ± 0.8 months old, $n = 14$). Data are presented as means $\pm$ s.d. Scale bars, 1 μ m.

aged 1–6 months, we calculated changes in total spine number as a function of age. These data were well fitted by a single-exponential equation (Fig. 4b; $70.0\% + 30.0\% \times (0.5)^{t/15.63 \text{ days}}$; $\chi^2 = 8.4$), indicating that after the first few months of postnatal development, the total number of spines remains relatively constant in adulthood^{7,8,21–23}. For a further quantitative assessment of changes in spine stability in the developing and adult animal, we calculated the average half-lives of spines in animals of different ages. We found that the decline in spine number in animals aged 1–2 months (Fig. 2f) was best fitted by a single-exponential equation ($68.1\% + 31.6\% \times (0.5)^{t/11.3 \text{ days}}$; $\chi^2 = 0.13$), suggesting that ~32% of spines are eliminated with a half-life of 11.3 days while the rest (68%) remain constant. In contrast, assuming that all spines are eliminated at a similar rate, spines in the adult animal (~4 months of age; Fig. 3m) would have an average half life of ~13.2 months ($100.7\% \times (0.5)^{t/396 \text{ days}}$; $\chi^2 = 6.7$). The switch from short to long spine half-life is well correlated with the end of the critical period in the mouse visual cortex¹⁰, indicating that spine stabilization might underlie the transition from a functionally plastic nervous system, during the critical period, to a less malleable one in adulthood.

Because experimental factors such as phototoxicity, anaesthesia, tissue injury, YFP expression and dendritic rotation between views would be more likely to induce artificial changes rather than long-term stability, we would have at most underestimated the degree of long-term spine stability. In addition, our observation of spine stability in mature mice but not in young ones suggests further that it is unlikely that these various factors induced artificial long-term stability selectively in adults. Because we obtained data primarily from spines located in cortical layers I and II, areas that have not traditionally been used to study synaptic plasticity, we do not know whether our observations of spine stability can be generalized to the entire dendritic tree of layer-5 pyramidal cells, different neuronal types, or neurons in regions other than the visual cortex. Further investigation with improved imaging approaches and new mouse

models will be necessary to address these questions. However, because the time course of synaptogenesis and synapse elimination have been shown to be similar in various regions of the brain^{8,21,23}, it is likely that other cortical areas share similar developmental regulation of spine stability. Indeed, preliminary observations over a 2-week interval suggest that spines are also remarkably stable in mature and less so in younger mouse barrel cortex (J.G. and W.-B.G., unpublished observations).

Although adult spines are stable over many months, it is not clear whether axonal terminals are stable or undergo constant turnover. In the peripheral nervous system, long-term *in vivo* imaging has revealed different degrees of synaptic structural stability and dynamism. For example, axon terminals in the young adult autonomic nervous system show gradual changes over periods of weeks and months^{26,27}. Yet mouse neuromuscular junctions²⁸ and submandibular ganglia (W.-B.G., G. Feng, J. R. Sanes and J. W. Lichtman, unpublished observations) maintain axonal branching patterns over many months in adulthood. Thus, as with spines, presynaptic axonal terminals might also demonstrate long-term stability in adulthood.

Our results show that dendritic spines in adult animals, past the critical period for the visual cortex, are remarkably stable with a half-life of more than 13 months (Supplementary Information). The gradual decline of spine plasticity extending into advanced postnatal development indicates that structural plasticity in young adulthood is significantly different from that in the more mature animal^{13,26,27}. The long-term stability of spines in the adult brain suggests that synapses can be maintained over the lifetime of an animal and could provide a structural basis for long-term information storage. In addition, the changes in spine length and diameter that we observed could provide a basis for short-term storage of information and could be involved in rapid plasticity, such as lesion-driven and experience-driven cortical remodeling^{5,6,11–13}. Such alterations in synaptic efficacy might ultimately regulate the formation and elimination of spines over longer periods and lead to long-term changes in neuronal circuitry. The extent and the timescale over which neuronal activity modulates adult spine stability require future studies. □

Methods

Transgenic mice

Transgenic mice (H-line¹⁴) were purchased from Jackson Laboratory and were housed and bred in Skirball Institute's animal facilities. All experiments were performed in accordance with animal protocols.

Surgical procedure for *in vivo* imaging

Mice aged 1–10 months were anaesthetized with an intraperitoneal injection (5.0 ml per kg body weight) containing 17 mg ml⁻¹ ketamine and 1.7 mg ml⁻¹ xylazine in 0.9% NaCl. The skull was exposed with a midline scalp incision and a region (~1 mm in diameter) over primary visual cortex was located based on stereotaxic coordinates. A high-speed drill (Fine Science Tools, Foster City, California, USA) was carefully used to reduce the skull thickness by ~50%. To avoid damaging (evidenced by dendritic blebbing, for example) the underlying cortex by friction-induced heat, a cool sterile solution was added to the skull periodically, and drilling was intermittent to permit heat dissipation. Skull thinning was completed by scraping the cranial surface with a micro-surgical blade (Surgistar no. 6400). For optimal image quality, skull thickness was reduced to a very thin layer (~30–50 µm).

To reduce respiration-induced movements, the skull was glued (ethyl cyanoacrylate; Elmer Products no. KG-585) to a stainless steel plate 400 µm thick with a central opening for skull access. The plate was screwed to two lateral bars located on both sides of the mouse head and fixed to a metal base. After imaging, the plate was detached from the skull, the scalp was sutured, and the animal was returned to the cage until the next viewing.

In vivo imaging of dendrites

The animal was inserted under a custom-made two-photon microscope similar to a previously described set-up²⁹. The Ti-sapphire laser was tuned to the excitation wavelength for YFP (920 nm) at a low laser power (<40 mW at the sample) to minimize the possibility of phototoxicity. A stack of image planes was acquired by using a water-immersion objective lens (60×, 0.9 numerical aperture; Olympus), an external detector and a digital zoom of 3.0–4.0×. The imaging depth was ~100–200 µm from the pial surface and the step size was 0.70 µm. Two-dimensional projections of three-dimensional image stacks containing dendritic segments of interest were used for all figures.

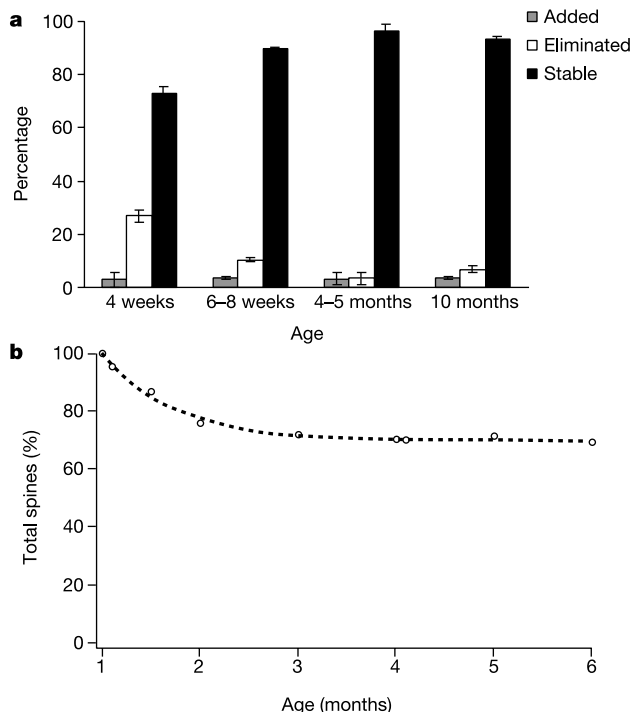


Figure 4 Spine stability increases with age. **a**, Spine stability, formation and elimination over 1-month intervals as a function of age. Data are presented as mean $\pm$ s.d. **b**, Percentage of total number of spines as a function of age. The data are fitted by a single-exponential equation (total percentage = $70.0\% + 30.0\% \times (0.5)^{t/15.63 \text{ days}}$, $\chi^2 = 8.4$).

Data quantification

The same dendritic segments (~5–20 μm in length; see Figs 2 and 3) were identified from three-dimensional image stacks taken from both views with high image quality (ratio of signal to background noise >4:1). These randomly chosen segments are probably mostly from different parent cells because of the high density of labelled neurons (>33 cells per 6,000 μm^2). The number and location of dendritic protrusions (protrusion length was more than one-third of dendritic shaft diameter) were identified in each view without prior knowledge of the animal's age, the interval between views or the order of the views. The total number of spines (N) was pooled from dendritic segments of different animals. Data throughout the text are presented as means $\pm$ s.d.

Filopodia were identified as long thin structures (generally larger than twice the average spine length, ratio of head diameter to neck diameter <1.2:1 and ratio of length to neck diameter >3:1). The remaining protrusions were classified as spines. Three-dimensional stacks were used to ensure that tissue movements and rotation between imaging intervals did not influence spine identification. Spines or filopodia were considered the same between two views on the basis of their spatial relationship to adjacent landmarks and their relative position to immediately adjacent spines (Figs 2 and 3). Spines in the second view were considered different if they were more than 0.7 μm from their expected positions based on the first view.

We chose 0.7 μm as a cut-off distance because apparent spine position can shift by up to ~0.3 μm in either direction along the axis of dendritic shafts owing to changes in spine morphology, slight tissue rotation and movements related to brain pulsation. We estimate the imaging resolution of our two-photon microscope (60 $\times$, numerical aperture 0.9, at 920 nm) to be ~0.7 μm . We found no significant changes in adult spine stability with three different cut-off distances (0.5, 0.7 and 0.9 μm), showing that our conclusion of spine stability is not dependent on minor changes in these criteria. Even though slight rotations of shafts and protrusions could either obscure existing spines (suggesting spine elimination) or reveal previously hidden spines (suggesting spine addition), such rotational artefacts would only underestimate the amount of stability observed.

For quantification of changes in spine morphology, we minimized the possibility of rotational artefacts by preselecting from three-dimensional image stacks only spines parallel to the imaging plane in both views. Dendrites containing saturated pixels were excluded. Imaging software (NIH ImageJ) was used to measure the spine length and head diameter of identical spines between views. The edges of spines were detected with a Sobel detector algorithm (ImageJ).

Curve fitting was done with the user-defined fitting function on Igor Pro (WaveMetrics, Lake Oswego, Oregon, USA). The single-exponential function was fitted to the data by using the built-in iterative Levenberg–Marquardt function to minimize χ^2 .

Received 28 August; accepted 4 November 2002; doi:10.1038/nature01276.

- Hubel, D. H., Wiesel, T. N. & LeVay, S. Plasticity of ocular dominance columns in monkey striate cortex. *Phil. Trans. R. Soc. Lond. B* **278**, 377–409 (1977).
- Rakic, P., Bourgeois, J. P. & Goldman-Rakic, P. S. Synaptic development of the cerebral cortex: implications for learning, memory, and mental illness. *Prog. Brain Res.* **102**, 227–243 (1994).
- Lichtman, J. W. & Colman, H. Synapse elimination and indelible memory. *Neuron* **25**, 269–278 (2000).
- Buonomano, D. V. & Merzenich, M. M. Cortical plasticity: from synapses to maps. *Annu. Rev. Neurosci.* **21**, 149–186 (1998).
- Florence, S. L., Taub, H. B. & Kaas, J. H. Large-scale sprouting of cortical connections after peripheral injury in adult macaque monkeys. *Science* **6**, 1117–1121 (1998).
- Jones, E. G. & Pons, T. P. Thalamic and brainstem contributions to large-scale plasticity of primate somatosensory cortex. *Science* **282**, 1121–1125 (1998).
- Lund, J. S., Boothe, R. G. & Lund, R. D. Development of neurons in the visual cortex (area 17) of the monkey (*Macaca nemestrina*): a Golgi study from fetal day 127 to postnatal maturity. *J. Comp. Neurol.* **176**, 149–188 (1977).
- Rakic, P., Bourgeois, J. P., Eckenhoff, M. F., Zecevic, N. & Goldman-Rakic, P. S. Concurrent overproduction of synapses in diverse regions of the primate cerebral cortex. *Science* **11**, 232–235 (1986).
- Shatz, C. J. & Stryker, M. P. Ocular dominance in layer IV of the cat's visual cortex and the effects of monocular deprivation. *J. Physiol. (Lond.)* **281**, 267–283 (1978).
- Antonini, A., Fagioli, M. & Stryker, M. P. Anatomical correlates of functional plasticity in mouse visual cortex. *J. Neurosci.* **1**, 4388–4406 (1999).
- Bao, S., Chan, V. T. & Merzenich, M. M. Cortical remodelling induced by activity of ventral tegmental dopamine neurons. *Nature* **412**, 79–83 (2001).
- Kleim, J. A. et al. Selective synaptic plasticity within the cerebellar cortex following complex motor skill learning. *Neurobiol. Learn. Mem.* **69**, 274–289 (1998).
- Knott, G. W., Quairiaux, C., Genoud, C. & Welker, E. Formation of dendritic spines with GABAergic synapses induced by whisker stimulation in adult mice. *Neuron* **34**, 265–273 (2002).
- Feng, G. et al. Imaging neuronal subsets in transgenic mice expressing multiple spectral variants of GFP. *Neuron* **28**, 41–51 (2000).
- Fiala, J. C., Feinberg, M., Popov, V. & Harris, K. M. Synaptogenesis via dendritic filopodia in developing hippocampal area CA1. *J. Neurosci.* **18**, 8900–8911 (1998).
- Dunaevsky, A., Blazeski, R., Yuste, R. & Mason, C. Spine motility with synaptic contact. *Nature Neurosci.* **4**, 685–686 (2001).
- Dailey, M. E. & Smith, S. J. The dynamics of dendritic structure in developing hippocampal slices. *J. Neurosci.* **16**, 2983–2994 (1996).
- Ziv, N. E. & Smith, S. J. Evidence for a role of dendritic filopodia in synaptogenesis and spine formation. *Neuron* **17**, 91–102 (1996).
- Dunaevsky, A., Tashiro, A., Majewska, A., Mason, C. & Yuste, R. Developmental regulation of spine motility in the mammalian central nervous system. *Proc. Natl Acad. Sci. USA* **96**, 13438–13443 (1999).
- Lendvai, B., Stern, E. A., Chen, B. & Svoboda, K. Experience-dependent plasticity of dendritic spines in the developing rat barrel cortex in vivo. *Nature* **404**, 876–881 (2000).
- Boothe, R. G., Greenough, W. T., Lund, J. S. & Wrege, K. A quantitative investigation of spine and

- dendrite development of neurons in visual cortex (area 17) of *Macaca nemestrina* monkeys. *J. Comp. Neurol.* **186**, 473–489 (1979).
- Huttenlocher, P. R. & de Courten, C. The development of synapses in striate cortex of man. *Hum. Neurobiol.* **6**, 1–9 (1987).
- Bourgeois, J. P., Goldman-Rakic, P. S. & Rakic, P. Synaptogenesis in the prefrontal cortex of rhesus monkeys. *Cereb. Cortex* **4**, 78–96 (1994).
- Harris, K. M. & Stevens, J. K. Dendritic spines of CA1 pyramidal cells in the rat hippocampus: serial electron microscopy with reference to their biophysical characteristics. *J. Neurosci.* **9**, 2982–2997 (1989).
- Murthy, V. N., Schikorski, T., Stevens, C. F. & Zhu, Y. Inactivity produces increases in neurotransmitter release and synapse size. *Neuron* **32**, 673–682 (2001).
- Purves, D. & Hadley, R. D. Changes in the dendritic branching of adult mammalian neurones revealed by repeated imaging *in situ*. *Nature* **315**, 404–406 (1985).
- Purves, D., Voyvodic, J. T., Magrassi, L. & Yawo, H. Nerve terminal remodeling visualized in living mice by repeated examination of the same neuron. *Science* **238**, 1122–1126 (1987).
- Lichtman, J. W., Magrassi, L. & Purves, D. Visualization of neuromuscular junctions over periods of several months in living mice. *J. Neurosci.* **7**, 1215–1222 (1987).
- Majewska, A., Yiu, G. & Yuste, R. A custom-made two-photon microscope and deconvolution system. *Pflügers Arch. Eur. J. Physiol.* **441**, 398–408 (2000).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank J. Lichtman, S. Burden and R. Yuste for critical comments on this manuscript. This work was supported by grants from the National Institutes of Health and the Ellison Foundation to W.-B.G. and by an Irene Diamond grant to M. L. Dustin for purchasing the imaging system.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to W.-B.G. (e-mail: gan@saturn.med.nyu.edu).

A role for casein kinase 2 α in the *Drosophila* circadian clock

Jui-Ming Lin^{*†}, Valerie L. Kilman^{*†}, Kevin Keegan^{*}, Brie Paddock^{*}, Myai Emery-Le[‡], Michael Rosbash[‡] & Ravi Allada^{*§}

^{*} Department of Neurobiology and Physiology, and [§] Department of Pathology, Northwestern University, Evanston, Illinois 60208, USA

[‡] Howard Hughes Medical Institute, Brandeis University, Waltham, Massachusetts 02454, USA

[†] These authors contributed equally to this work

Circadian clocks drive rhythmic behaviour in animals and are regulated by transcriptional feedback loops^{1,2}. For example, the *Drosophila* proteins Clock (Clk) and Cycle (Cyc) activate transcription of *period* (*per*) and *timeless* (*tim*). Per and Tim then associate, translocate to the nucleus, and repress the activity of Clk and Cyc. However, post-translational modifications are also critical to proper timing. Per and Tim undergo rhythmic changes in phosphorylation¹, and evidence supports roles for two kinases in this process: Doubletime (Dbt) phosphorylates Per^{3,4}, whereas Shaggy (Sgg) phosphorylates Tim⁵. Yet Sgg and Dbt often require a phosphoserine in their target site^{6,7}, and analysis of Per phosphorylation in *dbt* mutants^{3,8} suggests a role for other kinases. Here we show that the catalytic subunit of *Drosophila* casein kinase 2 (CK2 α) is expressed predominantly in the cytoplasm of key circadian pacemaker neurons. CK2 α mutant flies show lengthened circadian period, decreased CK2 activity, and delayed nuclear entry of Per. These effects are probably direct, as CK2 α specifically phosphorylates Per *in vitro*. We propose that CK2 is an evolutionary link between the divergent circadian systems of animals, plants and fungi.

To identify new components of circadian clocks, we screened

about 2,000 ethylmethane-sulphonate (EMS)-mutagenized *per^s* (short-period allele of *per*) lines for circadian behavioural defects. A dominant mutant, *Timekeeper* (*Tik*), was identified. *Tik* homozygotes do not live to adulthood. *Tik* heterozygotes exhibited behavioural rhythms approximately 1.5 h longer than *per^s* (Table 1). In a *per⁺* or *per^l* (long-period allele of *per*) background, *Tik* lengthened the period of the behavioural rhythm by 3 h, reflecting an allele-specific interaction between *per* and *Tik* (Table 1). During our studies, we identified a spontaneous partial revertant, *TikR* (Table 1), and we could not separate this revertant from *Tik* by recombination (0 out of 77 recombinants). *TikR* retains the flanking *Tik* markers, *roughoid*, *rosy* and *spineless*, and fails to complement the recessive lethality of *Tik*. The change in period in *Tik* mutants (about 3 h) exceeds that of nearly all heterozygous circadian mutants in *Drosophila*, suggesting that it might identify a protein of central importance in the circadian clock mechanism.

Using standard meiotic mapping and male recombination⁹, we mapped *Tik* to a region around the centromere of chromosome 3 (Table 2; see also Methods). We were unable to identify homozygous-viable *Tik* recombinants, nor do these recombinants complement each other for lethality, suggesting that *Tik* is a vital gene. We searched the non-recombinant interval for candidate genes, and initially focused on a candidate serine-threonine kinase, the catalytic α -subunit of casein kinase 2 (*CK2 α*)¹⁰. Sequencing of the *CK2 α* coding region from the wild-type parental strain and the initial *Tik* mutant identified two sequence changes, both resulting in coding changes: Met161Lys and Glu165Asp. The change from the non-charged Met 161 to the charged Lys occurs near the catalytic loop and within a hydrophobic binding pocket for ATP¹¹. We also sequenced the *TikR* mutant that we proposed was a new *Tik* revertant allele. In addition to the two original *Tik* coding changes, we identified an additional in-frame, 27-base-pair (bp) deletion coupled to an in-frame, 6-bp insertion, resulting in a deletion of 7 amino acids (234–240) and another amino acid change (Arg242-Glu). These mutations occur in residues that are highly conserved with their human counterpart.

To confirm that the observed mutations were functional, we established a biochemical assay of CK2 activity¹². CK2 was immunoprecipitated from adult head extracts using an antibody that recognizes the CK2 holoenzyme (two catalytic α - and two regulatory β -subunits)¹³. To verify that immunoprecipitated kinase activity was due to CK2, we measured the activity in the presence of heparin, a specific CK2 inhibitor. We found that 5 μ g ml⁻¹ heparin was able to reduce the activity of immunoprecipitated CK2 by about 90%, comparable to that of heparin against purified rat CK2 (data not shown). We did not observe any significant change in CK2 activity as a function of time in a 12-h light/dark cycle (Fig. 1a). We then measured CK2 activity in *CK2 α ^{Tik}* and *CK2 α ^{TikR}* heterozygotes. We found activity reductions in the *CK2 α ^{Tik}* heterozygotes and a return to near wild-type levels in *CK2 α ^{TikR}* (Fig. 1b). These *in vivo* biochemical measurements demonstrate that a loss (and its partial reversion) of CK2 enzymatic function is strongly correlated with changes in circadian period. Thus, the effects on the circadian

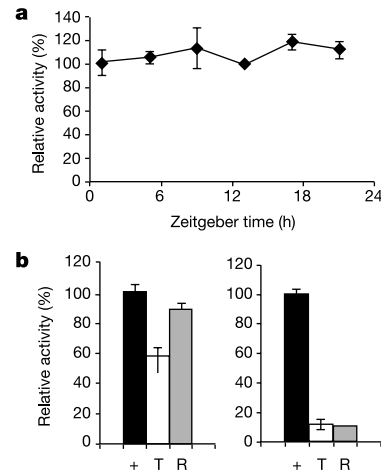


Figure 1 Biochemical activity of wild-type and mutant CK2 α . **a**, Diurnal regulation of CK2 activity. The graph displays the average of two experiments with ZT 13 set to 100. **b**, CK2 activity immunoprecipitated from head extracts (left panel). +, wild type; T, *Tik/+*; R, *TikR/+*. The graph shows the average of three experiments in duplicate with wild-type set to 100%. Error bars indicate standard error of the mean (s.e.m.). $P = 0.0003$ for wild type compared with *Tik/+*; $P = 0.003$ for *Tik/+* compared with *TikR/+*. The right panel shows recombinant CK2 α activity. Wild-type (+) is normalized to 100. Error bars indicate s.e.m. from six measurements (three experiments; $P < 0.00002$ for wild-type compared with *Tik/+*). Owing to the difficulty of obtaining material, *TikR* (R) activity was determined from one experiment.

cycle associated with the *CK2 α ^{Tik}* allele are probably due to a loss of CK2 function and not another effect of the mutation.

To determine clearly their biochemical defects, we assayed the activity of bacterially expressed wild-type and mutant CK2 α . We found that the recombinant CK2 α ^{Tik} mutant subunit was substantially less active than the wild type (Fig. 1b). We also found that the CK2 α ^{TikR} mutant had comparable catalytic activity to CK2 α ^{Tik}, suggesting that the reversion does not alter catalytic activity relative to CK2 α ^{Tik}. Notably, the CK2 α ^{TikR} protein was much less soluble, suggesting that folding of CK2 α ^{TikR} is altered *in vivo*, thus blocking incorporation into the CK2 holoenzyme. As a result, the CK2 holoenzyme in *CK2 α ^{TikR}* heterozygotes would consist largely of wild-type α -subunits. In this case, the period change (about 1 h) observed in these flies would be due to a reduction of wild-type gene dosage. Consistent with this hypothesis, homozygous *TikR* flies are not adult-viable.

If CK2 α is involved directly in circadian rhythms, it should be expressed in circadian pacemaker neurons. We performed immunohistochemistry on adult whole-mount dissected brains with an antiserum directed against a CK2 peptide sequence common to

Table 1 Circadian locomotor activity in *Tik* mutants

Genotype	Period (h) $\pm$ s.e.m.	Rhythmic (%)
+/+	23.6 $\pm$ 0.1	95 (43)
<i>Tik/+</i>	26.4 $\pm$ 0.1	88 (32)
<i>TikR/+</i>	24.3 $\pm$ 0.1	100 (24)
<i>per^s; +/+</i>	18.9 $\pm$ 0.1	100 (16)
<i>per^s;Tik/+</i>	20.4 $\pm$ 0.1	80 (15)
<i>per^l</i>	29.6 $\pm$ 0.1	100 (10)
<i>per^l;Tik/+</i>	32.5 $\pm$ 0.1	84 (7)

The number of flies analysed is indicated in parentheses in the final column. +, wild type; *Tik*, *Timekeeper*; *TikR*, *Timekeeper* revertant. Rhythmic (%) refers to the percentage of flies that were scored as rhythmic (see Methods). *per^s* and *per^l* are long and short period alleles of *period*.

Table 2 Mapping of *Tik*

Marker	Recombinants	Map
<i>cyc</i>	4/459 (R)	76C
<i>l(3)s2253</i>	2/112 (R)	77A
<i>ri</i>	3/230 (R)	77E
Ten-m05309	1/142 (R)*	79E1–79E2
nrmA37	4/393 (R)*	80A1–80A2
Centromere	–	80F–81F
EP3028	1/1,233 (L)*	82A?
<i>l(3)L1233</i>	1/896 (L)*	82B1–82B2
<i>Ki</i>	4/88 (L)	83D

The recombinants column indicates the number of recombinants out of the total number that were screened. R and L in parentheses indicate that *Tik* maps to the right or left, respectively. The position of EP3028 was inferred from genomic location. *cyc*, *cycle*; *ri*, *radius incompletus*; *Ki*, *Kinked*.

*Male recombination.

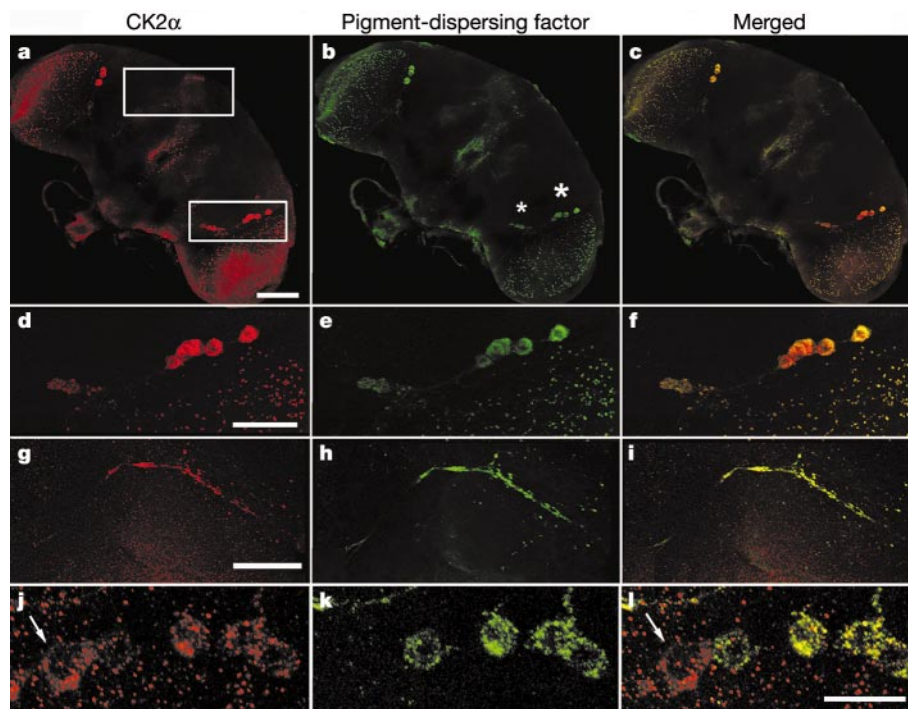


Figure 2 *Drosophila* CK2 α is primarily expressed in circadian pacemaker cells. **a–i**, Immunofluorescence of adult brain stained for CK2 α and Pigment-dispersing factor (Pdf) at ZT 18. The large and small asterisks indicate large and small lymph nodes, respectively. **d–f**, Magnified images of the lower boxed area in **a** showing lymph-node staining and punctate terminals of the large lymph nodes in the medulla.

g–i, Magnified images of the upper boxed area in **a**. The posterior brain shows double-labelled projections of the small lymph node to the superior protocerebrum. Scale bars: **a–c**, 100 μ m; **d–f**, 50 μ m; **g–i**, 50 μ m. **j–l**, Whole-mount immunofluorescence for CK2 α and Pdf in third instar larval brain. The white arrows indicate a CK2 α -positive but Pdf-negative cell body. Scale bar, 10 μ m.

both fly and mammalian CK2 α (anti-CK2 α). Anti-CK2 α specifically labels the cytoplasm and axonal projections of neurons in the lateral protocerebrum (Fig. 2a, d, g). To determine whether these neurons were the well-described pacemaker lateral neurons, we performed double labelling with an antiserum against Pigment-dispersing factor (Pdf)^{14–16}. This neuropeptide is specifically localized to small and large lateral neurons critical to behavioural rhythms (Fig. 2b, e, h). Consistent with its proposed circadian role, CK2 co-localizes with Pdf in these adult neurons (Fig. 2c, f, i).

The CK2 antiserum also variably labels two cells in the dorsal brain, but they are labelled much less intensely than the lateral neurons (data not shown). CK2 α is also probably expressed in the eyes, as CK2 activity is modestly reduced (about 20%) in *eyes absent* mutants (data not shown). Of note, CK2 α and Pdf also co-localize to neuronal termini, indicating that CK2 may regulate directly Pdf processing or release. CK2 α seems to be constitutively cytoplasmic as a function of the time of day (data not shown). To confirm that anti-CK2 α specifically recognizes CK2 α , we performed single

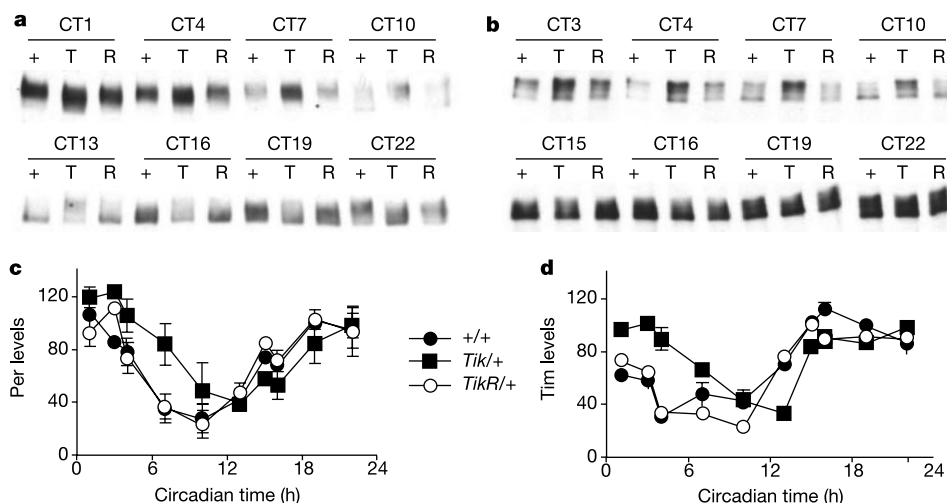


Figure 3 Circadian regulation of Per and Tim in CK2 α mutants. **a**, Western blot of Per. CT indicates circadian time, where CT0 is 12 h after lights off. +, wild type; T, *Tik*; R, *TikR*. **b**, Western blot of Tim. **c, d**, Quantification of Per and Tim western blots.

Each point is the average of two to three independent experiments except for CT3 and CT15 for both Per and Tim, and CT1 for Tim alone. Error bars indicate s.e.m.

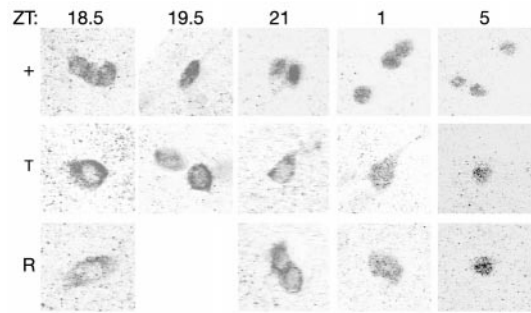


Figure 4 Nuclear entry of Per in homozygous *CK2α* mutants. +, wild type; T, *Tik/Tik* homozygotes; R, *TikR/TikR* homozygotes. Homozygous larvae were reared in a 12-h light/dark cycle. ZT, Zeitgeber time (ZT).

labelling with anti-CK2α alone and found a similarly specific expression pattern, which is specifically blocked by 10^{-6} M of recombinant CK2α but not by an unrelated antigen (data not shown). Western analysis confirmed this specificity. We also found that the CK2α antiserum labels a Pdf-negative cell in larval brains, indicating that the observed signal is not attributable to the Pdf signal (Fig. 2j–l). The tissue specificity of CK2α expression suggests that it has a specific function in circadian clocks.

To ascertain CK2 function on Per and Tim, we examined their levels and phosphorylation in heterozygotes. Flies were entrained in a light/dark cycle and were transferred to constant darkness (DD). During the early subjective day (circadian time, CT, 0–12) levels of Per in the *CK2α^{Tik}* mutants are increased and show delayed disappearance (Fig. 3a). Furthermore, there seems to be a modest increase in less-phosphorylated forms of Per. During the subjective night, Per phosphorylation and to a lesser extent accumulation are delayed relative to wild type. An increase in the level of Per and Tim in *CK2α^{Tik}* mutants is delayed by approximately 3 h relative to wild type, consistent with the period-lengthening effect of *CK2α^{Tik}* mutants (Fig. 3c and Table 1). Disappearance of Tim is also more strongly affected than accumulation in *CK2α^{Tik}* mutants (Fig. 3b, d). The Per and Tim profiles of *CK2α^{TikR}* heterozygotes are largely unchanged, consistent with its heterozygous behavioural and biochemical phenotypes.

To examine the phenotype of homozygous CK2α mutants, we performed immunofluorescence for Per in third-instar larval brains. Consistent with previous studies we found that Per staining in wild-type larvae is predominantly nuclear by Zeitgeber time (ZT) 21 (Fig. 4). In *CK2α^{Tik}* and *CK2α^{TikR}* homozygotes, however, Per nuclear entry is significantly delayed, with Per predominantly cytoplasmic at ZT21. By ZT1, the Per staining pattern in the *CK2α* mutants remains distinct from the wild-type nuclear pattern, appearing to be present in both cytoplasmic and nuclear compartments. These data demonstrate a strong effect of CK2α on Per nuclear entry and are consistent with the cytoplasmic expression of CK2α. Furthermore, we did not observe significant differences between *CK2α^{Tik}* and *CK2α^{TikR}*, consistent with biochemical data on recombinant proteins (Fig. 3b) and their recessive lethality. These data support our model that *CK2α^{TikR}* is a strong loss-of-function allele, lacking the strong dominant behavioural and biochemical effects of *CK2α^{Tik}*.

Previous studies have shown that CK2 phosphorylates Per and Tim *in vitro*, suggesting direct effects of CK2 (ref. 17). Consistent with previous reports, we found that bacterially expressed and purified CK2α can phosphorylate Per *in vitro* (Fig. 5a). Notably, this effect is specific, as we do not observe significant phosphorylation of the circadian transcription factor Cyc. We also found that Tim (amino acids 1–1159) is phosphorylated by CK2α to a lesser

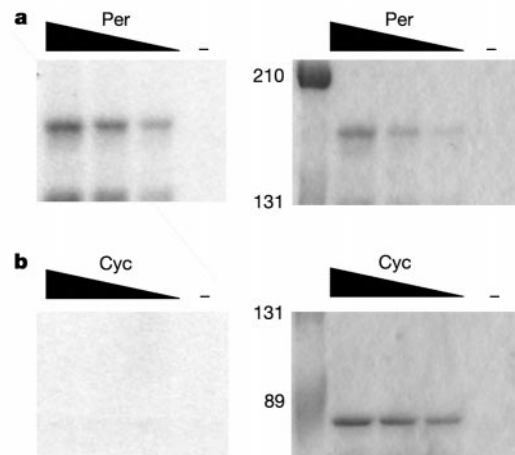


Figure 5 Phosphorylation by CK2α is period-specific. **a**, *In vitro* phosphorylation by CK2α of bacterially expressed and purified Per (left panel). The right panel shows Coomassie-stained Per as a loading control. Per varies from 0.125 to 0.5 μg. **b**, *In vitro* phosphorylation by CK2α of bacterially expressed and purified Cyc (left panel). The right panel shows Coomassie-stained Cyc as a loading control. Cyc varies from 0.25 to 0.5 μg. Numbers represent relative molecular mass ($\times 10^3$).

extent than Per (data not shown). These data are consistent with the direct regulation of Per and Tim by CK2.

Taken together, our analysis indicates a dedicated and direct role of CK2 in the *Drosophila* circadian clock mechanism. *CK2α^{Tik}* has one of the strongest phenotypes of any heterozygous circadian rhythm mutant. The modest nature of the biochemical defect further supports the hypothesis that the circadian clock is highly sensitive to CK2 activity. The CK2α expression pattern (Fig. 2) indicates that it has a specific role in circadian rhythms. Its localization to neuronal termini raises the possibility that CK2 links the clock to circadian outputs.

We favour the model that CK2 directly phosphorylates Per and Tim *in vivo*, promoting their transition to the nucleus. The evidence for this pattern is compelling for Per. Allele-specific interactions between *per* and *CK2α* alleles support the model of a direct interaction. It has been shown that CK2 can phosphorylate both Per and Tim *in vitro*¹⁷. By comparison with protein kinases A and C, CK2 is much more specific, as demonstrated by the paucity of other ³²P-labelled proteins in the immunoprecipitate¹⁷. We have also shown that CK2 can specifically phosphorylate Per *in vitro* (Fig. 5). Given the cytoplasmic expression of CK2α, this phosphorylation may serve as a signal for nuclear entry and subsequent degradation, explaining the delayed nuclear entry and disappearance of these proteins in *CK2α^{Tik}* mutants.

This study also establishes an evolutionary connection between animal, plant and fungal circadian systems—genetic studies have revealed clock components in plants (*Arabidopsis*) and fungi (*Neurospora*). These studies suggest that clocks have arisen several times in evolution¹⁸; however, studies in both *Arabidopsis* and *Neurospora* have linked CK2 to circadian timekeeping^{19–21}. CK2 is therefore a gene involved in circadian rhythms that is shared between all three phylogenetic kingdoms. Indeed, the fly enzyme (amino acids 7–322) shares 77% and 72% identity with the *Arabidopsis* and *Neurospora* enzymes, respectively. We propose that the conserved role for CK2 is driven by the need to avoid mutagenic ultraviolet light. CK2 has a pivotal role in the response to ultraviolet radiation from yeast to humans^{22,23}. Consistent with this model, cryptochromes, components of plant and animal circadian systems, are homologous to ultraviolet-dependent DNA repair enzymes²⁴. Given this strong conservation, the circadian role of mammalian CK2α

homologues should be of great interest.

Note added in proof: Recent work from F. R. Jackson and others demonstrates a role for the β -subunit of CK2 in *Drosophila* circadian rhythms. □

Methods

Drosophila genetics

per^s; *ry⁵⁰⁶* flies were used for mutagenesis and as controls for identifying mutations²⁵. Fly stocks are from Bloomington Stock Center. Initial mapping with *roughoid* (*ru*), *spineless* (*ss*) and *claret* (*ca*) identified approximately 10% recombination between *ry⁵⁰⁶* and *Tik*. Scoring of flanking markers indicated that *Tik* mapped to the left of *ry*. Further mapping of informative recombinants is summarized in Table 2. Uninformative meiotic recombination experiments are as follows: Y257 (79C; 0 out of 313), I(3)L7251 (79F; 0 out of 211), I(3)J1E6 (82A; 0 out of 111). We were unable to find deficiencies that failed to complement the circadian or lethal phenotypes.

Circadian locomotor activity analysis

Flies were entrained for 2–5 days in 12-h light/dark conditions, and locomotor activity behaviour was monitored subsequently in constant darkness for 6–8 days as described previously using Trikinetics locomotor activity monitors²⁶. We performed χ^2 periodogram analysis ($\alpha = 0.01$) of locomotor activity using the ClockLab software package (Actimetrics). Flies whose χ^2 statistic exceeded the significance line by 5 were judged to be significantly rhythmic, and were used for period calculations.

Bacterial expression and recombinant protein purification

Wild-type and mutant *CK2 α* coding regions were amplified (Pfu turbo; Stratagene) from genomic DNA, subcloned into pET-30(a), expressed in BL21, and purified using His and S tags (Novagen). We confirmed the presence or absence of mutations by sequencing. Per was amplified by polymerase chain reaction (PCR) (Pfu turbo; Stratagene), subcloned into pQE-80L (Qiagen), expressed in DH5 α , and purified using a His tag, following the manufacturer's instructions. His-tagged Per adds a further 32 amino acids to the amino-terminus, which does not contain any consensus CK2 sites. Per phosphorylation results were confirmed by testing two independent clones. Expression of glutathione S-transferase (GST)–Tim (1–1159) and GST–Cyc (also known as dBmal) was performed as in ref. 5.

CK2 activity assays

Fly head extracts were prepared⁸ and normalized to the same protein concentration by Bradford assay. Immunoprecipitation was performed¹⁷ with a CK2 antibody that recognizes the *Drosophila* holoenzyme (anti-CK2H)¹³, with the following changes: 40 μ l beads, 500 μ l extraction buffer washes, 0.5 μ l anti-CK2, and 150 μ l of extract was added (5–8 mg ml^{−1}). The mixture was incubated end-over-end for 4 h. Beads were washed twice for 5 min then once with 500 μ l CK2 assay buffer (50 mM Tris, pH 8.5, 100 mM NaCl, 10 mM MgCl₂). Beads were then incubated with 25 μ l CKII activity assay buffer, 1 mg ml^{−1} biotinylated peptide containing an optimized CK2 target site, [biotin]-RRREETEEE-[OH] (Princeton BioMolecules), 0.1 mM ATP, 0.6 μ l [γ -³²P]ATP. This reaction was incubated end-over-end at room temperature for 10 min. For recombinant CK2 α , 25 μ l CK2 assay buffer (50 mM Tris, pH 8.5, 100 mM NaCl, 10 mM MgCl₂, 1 mg ml^{−1} biotinylated peptide, 1.5 μ l [γ -³²P]ATP) was incubated for 30 min at room temperature with 2.5 μ g recombinant protein. We added 15 μ l of 7.5 M guanidine-HCl to stop the reaction. A total of 25 μ l supernatant from the reaction was transferred to a SAM² biotin-capture membrane (Promega) until dry. Each membrane was washed individually using a 6-well culture plate at room temperature as follows: 5 ml of 1 M NaCl four times for 1 min; 5 ml of 1 M NaCl with 1% phosphoric acid twice for 1 min; 5 ml of water once for 1 min; and 95% alcohol once for 15 s.

For Per, Tim and Cyc *in vitro* kinase reactions, 100 μ l CK2 kinase assay buffer (50 mM Tris, pH 8.5, 100 mM NaCl, 10 mM MgCl₂, 0.5 μ l [γ -³²P]ATP) containing recombinant protein was incubated at room temperature for 10 min. We used 0.2 μ g CK2 α for each assay, and we used 0.1–0.5 μ g of Per and Cyc. After the kinase reaction, the proteins were precipitated by tricarboxylic acid to remove free nucleotide, and boiled in $\times$ 2 SDS gel loading buffer.

Immunohistochemistry

Adult and larval brains were fixed and stained for whole-mount immunofluorescence²⁷ with the following exceptions: 10% donkey normal serum and no bovine serum albumin were used for blocking and antibody incubations, primary antibody incubations were $\geq$ 24 h, and only secondary (not tertiary) antibodies were used. Affinity purified antibodies were diluted to 1.7–8.5 μ g ml^{−1} (rabbit anti-CK2 α (StressGen); 1:100–200 rat anti-Pd²⁸, 1:100–300 Cy3-labelled donkey anti-rabbit, and/or 1:100–300 fluorescein isothiocyanate-donkey anti-rat (Jackson ImmunoResearch)). Rabbit anti-Per antiserum was diluted 1:4,000 after pre-incubation with *per⁰* embryos. Tissues were mounted in Fluoromount-G (Southern Biotechnology) or 80% glycerol in PBS. Preparations were imaged on a Leica laser scanning confocal microscope (optical sections every 1–1.5 μ m) and images were processed with Adobe Photoshop. Filter settings and laser strengths were chosen to minimize detection of green emission in the red channel in double labels. Also, for some preparations, the two channels were collected separately, using only the appropriate excitation wavelength to avoid cross-excitation. All results were qualitatively verified with single-labelling experiments.

Western blotting

Quantification of western blotting⁸ was performed on scanned autoradiographs and densitometry was performed using Scion Image. Equal loading and transfer was confirmed using Ponceau-S staining of membranes.

Received 30 July; accepted 21 October 2002; doi:10.1038/nature01235.

Published online 24 November 2002.

- Allada, R., Emery, P., Takahashi, J. S. & Rosbash, M. Stopping time: the genetics of fly and mouse circadian clocks. *Annu. Rev. Neurosci.* **24**, 1091–1119 (2001).
- Toh, K. L. *et al.* An hPer2 phosphorylation site mutation in familial advanced sleep phase syndrome. *Science* **291**, 1040–1043 (2001).
- Price, J. L. *et al.* double-time is a novel *Drosophila* clock gene that regulates PERIOD protein accumulation. *Cell* **94**, 83–95 (1998).
- Kloss, B. *et al.* The *Drosophila* clock gene double-time encodes a protein closely related to human casein kinase I ϵ . *Cell* **94**, 97–107 (1998).
- Martinek, S., Inonog, S., Manoukian, A. S. & Young, M. W. A role for the segment polarity gene shaggy/GSK-3 in the *Drosophila* circadian clock. *Cell* **105**, 769–779 (2001).
- Fiol, C. J., Wang, A., Roeske, R. W. & Roach, P. J. Ordered multisite protein phosphorylation. Analysis of glycogen synthase kinase 3 action using model peptide substrates. *J. Biol. Chem.* **265**, 6061–6065 (1990).
- Flotow, H. *et al.* Phosphate groups as substrate determinants for casein kinase I action. *J. Biol. Chem.* **265**, 14264–14269 (1990).
- Suri, V., Hall, J. C. & Rosbash, M. Two novel doubletime mutants alter circadian properties and eliminate the delay between RNA and protein in *Drosophila*. *J. Neurosci.* **20**, 7547–7555 (2000).
- Chen, B., Chu, T., Harms, E., Gergen, J. P. & Strickland, S. Mapping of *Drosophila* mutations using site-specific male recombination. *Genetics* **149**, 157–163 (1998).
- Saxena, A., Padmanabha, R. & Glover, C. V. Isolation and sequencing of cDNA clones encoding alpha and beta subunits of *Drosophila melanogaster* casein kinase II. *Mol. Cell Biol.* **7**, 3409–3417 (1987).
- Niefend, K., Guerra, B., Pinna, L. A., Issinger, O. G. & Schomburg, D. Crystal structure of the catalytic subunit of protein kinase CK2 from *Zea mays* at 2.1 Å resolution. *EMBO J.* **17**, 2451–2462 (1998).
- Goueli, B. S., Hsiao, K., Tereba, A. & Goueli, S. A. A novel and simple method to assay the activity of individual protein kinases in a crude tissue extract. *Anal. Biochem.* **225**, 10–17 (1995).
- Dahmus, G. K., Glover, C. V., Brutlag, D. L. & Dahmus, M. E. Similarities in structure and function of calf thymus and *Drosophila* casein kinase II. *J. Biol. Chem.* **259**, 9001–9006 (1984).
- Helfrich-Forster, C. The period clock gene is expressed in central nervous system neurons which also produce a neuropeptide that reveals the projections of circadian pacemaker cells within the brain of *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **92**, 612–616 (1995).
- Park, J. H. & Hall, J. C. Isolation and chronobiological analysis of a neuropeptide pigment-dispersing factor gene in *Drosophila melanogaster*. *J. Biol. Rhythms* **13**, 219–228 (1998).
- Renn, S. C., Park, J. H., Rosbash, M., Hall, J. C. & Taghert, P. H. A pdf neuropeptide gene mutation and ablation of PDF neurons each cause severe abnormalities of behavioral circadian rhythms in *Drosophila*. *Cell* **99**, 791–802 (1999); erratum **101**, 113 (2000).
- Zeng, H., Qian, Z., Myers, M. P. & Rosbash, M. A light-entrainment mechanism for the *Drosophila* circadian clock. *Nature* **380**, 129–135 (1996).
- Harmer, S. L., Panda, S. & Kay, S. A. Molecular bases of circadian rhythms. *Annu. Rev. Cell Dev. Biol.* **17**, 215–253 (2001).
- Sugano, S., Andronis, C., Green, R. M., Wang, Z. Y. & Tobin, E. M. Protein kinase CK2 interacts with and phosphorylates the *Arabidopsis* circadian clock-associated 1 protein. *Proc. Natl Acad. Sci. USA* **95**, 11020–11025 (1998).
- Sugano, S., Andronis, C., Ong, M. S., Green, R. M. & Tobin, E. M. The protein kinase CK2 is involved in regulation of circadian rhythms in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **96**, 12362–12366 (1999).
- Yang, Y., Cheng, P. & Liu, Y. Regulation of the *Neurospora* circadian clock by casein kinase II. *Genes Dev.* **16**, 994–1006 (2002).
- Ghavidel, A. & Schultz, M. C. TATA binding protein-associated CK2 transduces DNA damage signals to the RNA polymerase III transcriptional machinery. *Cell* **106**, 575–584 (2001).
- Keller, D. M. *et al.* A DNA damage-induced p53 serine 392 kinase complex contains CK2, hSpt16, and SSRP1. *Mol. Cell* **7**, 283–292 (2001).
- Cashmore, A. R., Jarillo, J. A., Wu, Y. J. & Liu, D. Cryptochromes: blue light receptors for plants and animals. *Science* **284**, 760–765 (1999).
- Allada, R., White, N. E., So, W. V., Hall, J. C. & Rosbash, M. A mutant *Drosophila* homolog of mammalian Clock disrupts circadian rhythms and transcription of period and timeless. *Cell* **93**, 791–804 (1998).
- Hamblen, M. *et al.* Germ-line transformation involving DNA from the period locus in *Drosophila melanogaster*: overlapping genomic fragments that restore circadian and ultradian rhythmicity to *per⁰* and *per⁺* mutants. *J. Neurogenet.* **3**, 249–291 (1986).
- Kaneko, M., Helfrich-Forster, C. & Hall, J. C. Spatial and temporal expression of the period and timeless genes in the developing nervous system of *Drosophila*: newly identified pacemaker candidates and novel features of clock gene product cycling. *J. Neurosci.* **17**, 6745–6760 (1997).
- Park, J. H. *et al.* Differential regulation of circadian pacemaker output by separate clock genes in *Drosophila*. *Proc. Natl Acad. Sci. USA* **97**, 3608–3613 (2000).

Acknowledgements We thank J. Rutla for assistance and guidance in conducting the genetic screen; J. Biswas for assistance with sequencing; A. McElvaine for cloning; L. McCarty for western blots; R. Scharnweber for helping to set up the CK2 activity assays; Bloomington Stock Center for fly stocks; C. Glover for CK2 antibodies; and the Northwestern University Biological Imaging Facility for assistance with confocal microscopy.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.A. (e-mail: r-allada@northwestern.edu).

Glucose transporter recycling in response to insulin is facilitated by myosin Myo1c

Avirup Bose, Adilson Guilherme, Stacey I. Robida, Sarah M. C. Nicoloso, Qiong L. Zhou, Zhen Y. Jiang, Darcy P. Pomerleau & Michael P. Czech

Program in Molecular Medicine, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA

Insulin stimulates glucose uptake in muscle and adipocytes by signalling the translocation of GLUT4 glucose transporters from intracellular membranes to the cell surface^{1,2}. The translocation of GLUT4 may involve signalling pathways that are both independent of and dependent on phosphatidylinositol-3-OH kinase (PI(3)K)^{3–5}. This translocation also requires the actin cytoskeleton^{6–8}, and the rapid movement of GLUT4 along linear tracks may be mediated by molecular motors⁹. Here we report that the unconventional myosin Myo1c is present in GLUT4-containing vesicles purified from 3T3-L1 adipocytes. Myo1c, which contains a motor domain, three IQ motifs and a carboxy-terminal cargo domain, is highly expressed in primary and cultured adipocytes. Insulin enhances the localization of Myo1c with GLUT4 in cortical tubulovesicular structures associated with actin filaments, and this colocalization is insensitive to wortmannin. Insulin-stimulated translocation of GLUT4 to the adipocyte plasma membrane is augmented by the expression of wild-type Myo1c and inhibited by a dominant-negative cargo domain of Myo1c. A decrease in the expression of endogenous Myo1c mediated by small interfering RNAs inhibits insulin-stimulated uptake of 2-deoxyglucose. Thus, myosin Myo1c functions in a PI(3)K-independent insulin signalling pathway that controls the movement of intracellular GLUT4-containing vesicles to the plasma membrane.

We used a method developed in our laboratory¹⁰ to identify proteins present in GLUT4-enriched membranes purified from differentiated 3T3-L1 adipocytes. Membrane fractions from the final equilibrium density gradient (10–65%) containing the insulin-sensitive pool of GLUT4 were combined and collected by centrifugation, resolved by a 5–15% gradient on SDS polyacrylamide gel electrophoresis (SDS–PAGE) gels and visualized with silver stain (Supplementary Fig. S1). Bands were excised and subjected to proteolytic digestion before analysis by mass spectrometry. We identified several proteins, including sortilin I (ref. 11), SCAMP¹², IRAP¹³ and VAMP2 (ref. 14), that have been reported to be present in GLUT4 vesicles. In addition, three peptides derived from a protein band with a relative molecular mass of 118,000 (M_r 118K) were found to correspond to amino-acid sequences of the class I unconventional myosin Myo1c (previously called MM1 β /Myr2)¹⁵.

Myo1c contains an amino-terminal catalytic motor domain, three centrally located IQ motifs and a unique C-terminal domain thought to be involved in binding to cargo receptor proteins (refs 16, 17, and Fig. 1a). This cargo domain contains several proline-rich domains that may bind Src homology domain 3 (SH3) domains. Other members of this family include Myo1a, Myo1b (Fig. 1a) and Myo1d–Myo1h, some of which have been implicated in vesicle and organelle transport¹⁸, recycling of endosomal and Golgi membranes^{19,20} and mechanochemical regulation of calcium channels in hair cells²¹.

We also identified Myo1c in an independent screen for highly expressed transcripts in adipocytes using gene arrays. This was confirmed by northern blot analysis, which showed high expression of Myo1c in primary adipocytes, and a large increase in expression on differentiation of 3T3-L1 fibroblasts to adipocytes (Fig. 1b). We

compared this expression profile to that of the related Myo1b isoform (Fig. 1a), which has been proposed to function in transferrin endocytosis²². Myo1b is virtually absent from adipocytes, although both myosins are relatively highly expressed in lung (Fig. 1b). Western blotting using an antibody specific to Myo1c that we raised in rabbits validated the predominant expression of Myo1c in mouse lung and fat, and also showed that there is much lower but significant expression in insulin-sensitive skeletal and heart muscle (Fig. 1c). We confirmed that expression of the dominant-negative cargo domain of Myo1b in COS-1 cells disrupts internalization of transferrin to perinuclear membranes²², but the divergent cargo domain of Myo1c did not (Supplementary Fig. S2).

Endogenous Myo1c is localized mostly in the cytoplasm of cultured adipocytes, in a granular or punctate pattern that is similar to that reported in other types of cell²³. Myo1c could also be seen near the plasma membrane in about a third of the cells (Fig. 1d, e). Insulin treatment of 3T3-L1 adipocytes increased by about twofold the number of adipocytes in which Myo1c could be visualized in this cortical area (Fig. 1d, e, arrowheads). This translocation of Myo1c coincided with the augmented formation of cortical F-actin visualized by staining with rhodamine–phalloidin (ref. 8 and Supplementary Fig. S3) and GLUT4 in the subplasma-membrane region (Supplementary Fig. S4). Studies have suggested that two

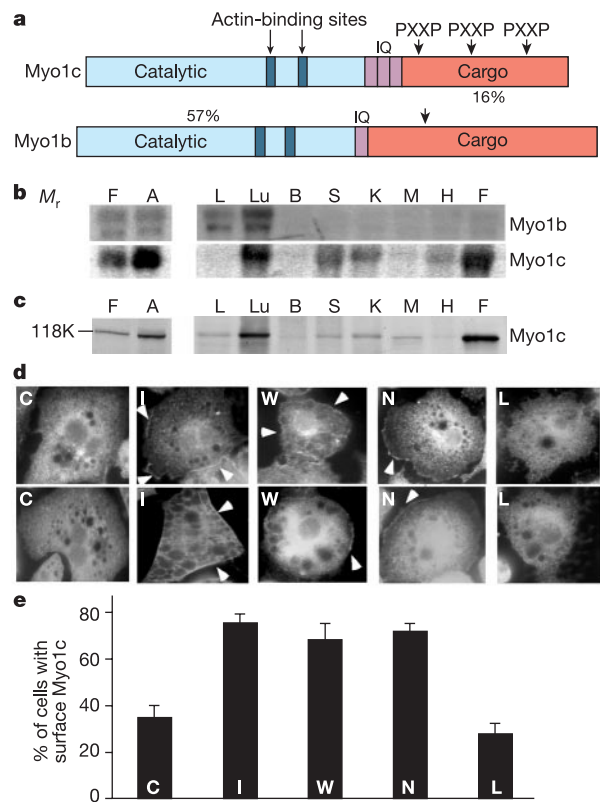


Figure 1 Insulin augments cortical Myo1c localization in 3T3-L1 adipocytes.

a, Comparison of the domains of mouse Myo1c and Myo1b. Percentages reflect sequence identities in the motor and cargo domains. **b**, Myo1c and Myo1b mRNA levels in 3T3-L1 fibroblasts (F), differentiated adipocytes (A) and in mouse liver (L), lungs (Lu), brain (B), spleen (S), kidney (K), skeletal muscle (M), heart (H) and fat (F). **c**, Quantities of Myo1c protein in the same tissues. **d**, Differentiated 3T3-L1 adipocytes were serum-starved (C) and then either treated with insulin (I) or treated with wortmannin (W), nocodazole (N) or latrunculin B (L) before treatment with insulin. The cells were then fixed and stained with anti-Myo1c. Images were analysed using digital microscopy, and a representative field from three separate experiments is shown. **e**, About 100 cells from the treatments described in **d** were scored blindly for plasma membrane staining of Myo1c. Cells in which Myo1c was present in any part of the plasma membrane were scored as positive.

insulin signalling pathways may be required to stimulate GLUT4 recycling to the plasma membrane: one dependent on functional p85/p110-type PI(3)K²⁴, and one independent of this kinase^{3–5}.

We previously reported that the role of a PI(3)K-independent insulin signalling pathway is to cause polymerization of cortical actin through the actin-regulator N-WASP⁸. We found that 100 nM wortmannin did not attenuate the ability of insulin to elicit acute Myo1c movement to the cell periphery (Fig. 1d, e), whereas the drug completely inhibited PI(3)K activity and GLUT4 translocation under the same experimental conditions (data not shown). Treatment of 3T3-L1 adipocytes with the actin-depolymerizing agent latrunculin B inhibited insulin-stimulated translocation of Myo1c to the cell periphery (Fig. 1d, e), whereas nocodazole, which we could confirm disrupts the microtubule network in adipocytes²⁵, did not.

Analysis of this peripheral region of insulin-treated 3T3-L1 adipocytes by electron microscopy showed a complex network of apparent tubulo-vesicular membranous structures (Fig. 2, see also Supplementary Fig. S5). Antibodies to both Myo1c and GLUT4 showed extensive binding to these structures in close proximity. Most of this filamentous network of intracellular membranes at the cell periphery disappeared when these cells were treated with the

actin-depolymerizing agent latrunculin B (data not shown). The colocalization of GLUT4 and Myo1c in these actin-based membranes was very similar to the model developed for spreading melanin granules in pigment cells, in which unconventional myosin Va is colocalized with F-actin at dendritic destinations of these granules²⁶. Capture of the melanosomes by myosin Va at F-actin structures seems to be the predominant mode of peripheral accumulation of these organelles²⁷.

The idea that Myo1c may function similarly to drive GLUT4 motility near the plasma membrane is consistent with studies by our laboratory^{8,28} and others^{6,7} that have shown that the actin cytoskeleton is required for GLUT4 translocation by insulin. In addition, in live 3T3-L1 adipocytes linear movements of green fluorescent protein (GFP)-tagged GLUT4 along apparent tracks towards and near the cell periphery have been documented⁹. The results shown in Figs 1 and 2 indicate that Myo1c may be a target for an insulin signalling pathway that augments the localization of both F-actin and this molecular motor to sites where it can facilitate the movement of GLUT4.

We used three distinct approaches to test whether Myo1c functions in insulin-stimulated GLUT4 movements to the cell surface. First, a haemagglutinin A (HA)-tagged dominant inhibitory cargo domain of Myo1c, HA-Myo1c(T), was co-transfected into cultured adipocytes with a plasmid encoding GLUT4 fused at its C terminus to enhanced GFP (EGFP) and tagged with a Myc epitope in its first exofacial loop⁸. When 3T3-L1 adipocytes expressing only Myc-GLUT4-EGFP were stimulated with 100 nM insulin, fixed

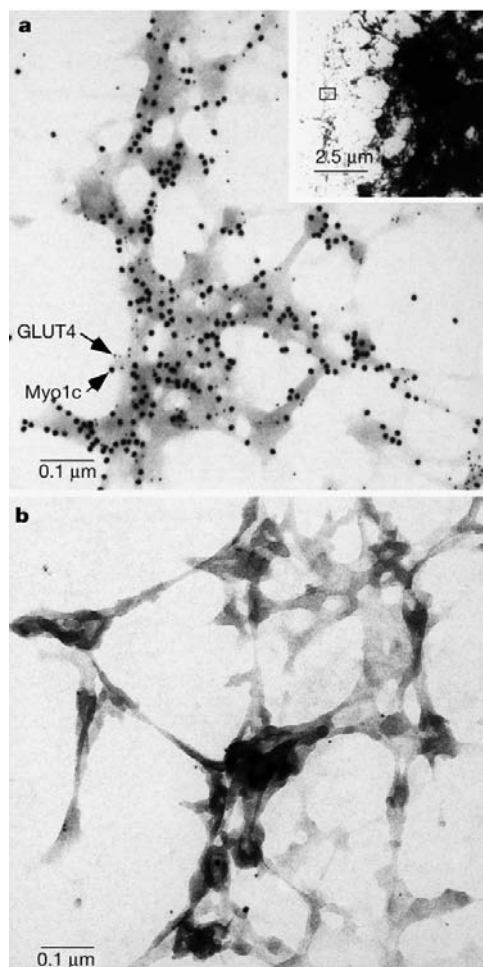


Figure 2 Whole-mount electron microscopy of the peripheral region of an insulin-stimulated 3T3-L1 adipocyte. Differentiated 3T3-L1 adipocytes were grown on Formvar-coated gold grids. **a**, Inset, low magnification ($\times 4,000$) of a part of an insulin-treated 3T3-L1 adipocyte. The dark area indicates the nucleus of the cell. Main image, high magnification ($\times 100,000$) of the boxed area in the inset shows a region near the edge of that cell. The bigger particles (12 nm) represent Myo1c, the smaller particles (6 nm) represent GLUT4. **b**, High magnification ($\times 100,000$) image of a similar cortical region of an adipocyte labelled with goat and rabbit IgGs. Images are representative of three separate experiments.

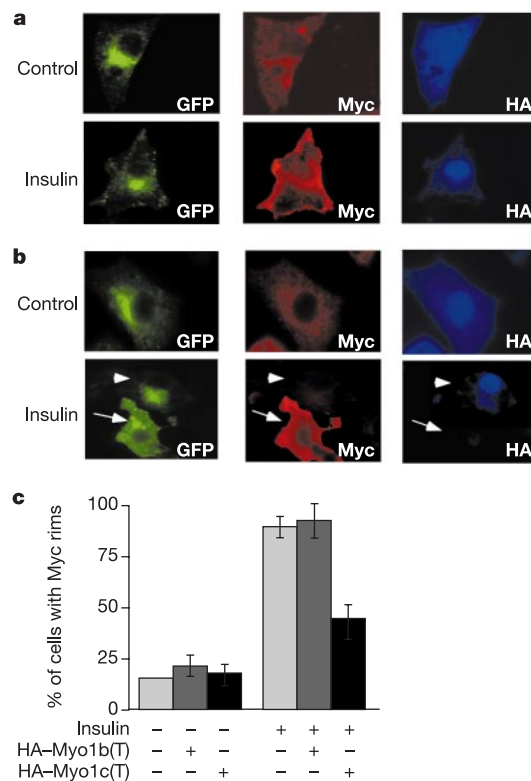


Figure 3 Dominant-negative cargo domain of Myo1c but not Myo1b inhibits insulin-stimulated Myc-GLUT4-GFP translocation to the plasma membrane. **a**, **b**, Differentiated 3T3-L1 adipocytes were co-transfected with Myc-GLUT4-GFP and either HA-Myo1b(T) (**a**) or HA-Myo1c(T) (**b**). Arrow in **b** indicates a cell expressing only Myc-GLUT4-GFP, arrowhead represents a cell expressing both Myc-GLUT4-GFP and HA-Myo1c(T). **c**, Number of transfected cells showing cell-surface anti-Myc staining under control and insulin-stimulated conditions. Two hundred cells from four separate experiments were scored blindly for anti-Myc staining throughout the plasma membrane under control and insulin-stimulated conditions. Error bars indicate standard deviation.

without permeabilization, and probed with antibodies against Myc, intense antibody-stained cell rims could be detected in almost 90% of the cells, as compared with only 15% of unstimulated cells (Fig. 3c). Adipocytes expressing the Myo1c cargo domain HA-Myo1c(T) along with Myc-GLUT4-EGFP showed a markedly decreased insulin effect, with anti-Myc signal at the cell rims detected on only 48% of the cells (Fig. 3b, c). By contrast, the HA-tagged cargo domain of Myo1b, HA-Myo1b(T), which acts as a dominant-negative inhibitor of transferrin endocytosis (ref. 22 and Supplementary Fig. S2), had no detectable effect on the insulin-stimulated translocation of Myc-GLUT4-GFP to the plasma membrane (Fig. 3a-c).

Second, we transfected small interfering RNAs (siRNAs)^{29,30} into cultured adipocytes to induce specific degradation of the Myo1c messenger RNA in these cells. Twenty four hours after transfection, the amounts of Myo1c protein decreased by about 40% in cells transfected with a combination of two siRNA species and about 25% in cells transfected with one siRNA species (Fig. 4a, b). Expression of two other control proteins, myosin 2b and EHD2, was unaffected (Fig. 4a). Insulin-stimulated 2-deoxyglucose uptake in adipocytes transfected with either siRNA 6 or both siRNA 6 and siRNA 2 was significantly inhibited as compared with cells transfected with scrambled siRNA at concentrations of both 1 nM and 100 nM insulin. These results indicate that siRNA-mediated attenuation of Myo1c expression causes a corresponding decrease in GLUT4 responsiveness to insulin.

Third, we expressed the full-length Myo1c protein, tagged with yellow fluorescent protein (YFP), in differentiated adipocytes also expressing Myc-GLUT4 fused to a cyan fluorescent protein (CFP) tag, and analysed the expression of Myc-GLUT4-CFP at the plasma membrane. Adipocytes expressing high quantities of YFP-Myo1c showed a significant enhancement of Myc-GLUT4-CFP on the cell surface under insulin-stimulated conditions (Fig. 5a). Analysis of the intensity of cell-surface anti-Myc staining under these conditions, normalized to the expression of Myc-GLUT4-CFP in each cell, confirmed a significant increase in cell-surface Myc-GLUT4-CFP in both unstimulated and in insulin-stimulated cells expressing

high quantities of Myo1c-YFP (Fig. 5a, b). Thus, attenuating Myo1c function by a dominant-negative construct (Fig. 3) or by siRNA-mediated Myo1c mRNA degradation (Fig. 4) inhibits insulin action on the glucose transporter, whereas increasing Myo1c expression augments translocation of GLUT4 (Fig. 5).

Our findings suggest the following hypothesis: insulin signalling through a PI(3)K-independent mechanism augments the polymerization of actin filaments at the cell periphery⁸, which in turn recruit the Myo1c motor protein by binding to its actin-binding sites. These events occur in parallel with PI(3)K-dependent steps required for GLUT4 translocation. Myo1c bound to actin filaments associates with actin-based tubulo-vesicular membranes containing GLUT4 (Fig. 2). Myo1c function in this process seems to drive the motility of GLUT4-containing membranes in the subplasma-membrane region before their docking and fusion, although these latter processes might also be facilitated by motor activity. □

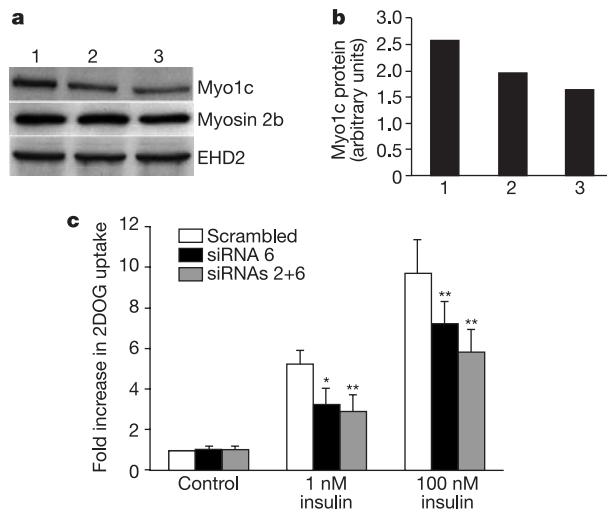


Figure 4 Inhibition of insulin-stimulated 2-deoxyglucose uptake in differentiated 3T3-L1 adipocytes by siRNA-mediated degradation of Myo1c mRNA. **a**, Amounts of Myo1c, myosin 2b and EHD2 protein in differentiated 3T3-L1 adipocytes after expression of scrambled siRNA (lane 1), siRNA 6 (lane 2) or siRNAs 6 and 2 (lane 3). Data are representative of five separate experiments. **b**, Quantification of the Myo1c protein shown in **a** using a scanning densitometer. **c**, Insulin-stimulated [³H]-2-deoxyglucose (2DOG) uptake in differentiated 3T3-L1 adipocytes transfected with either scrambled siRNA, siRNA 6, or siRNAs 6 and 2. Results are the average of three separate experiments. Asterisk, $P < 0.0004$, double asterisk, $P < 0.0001$.

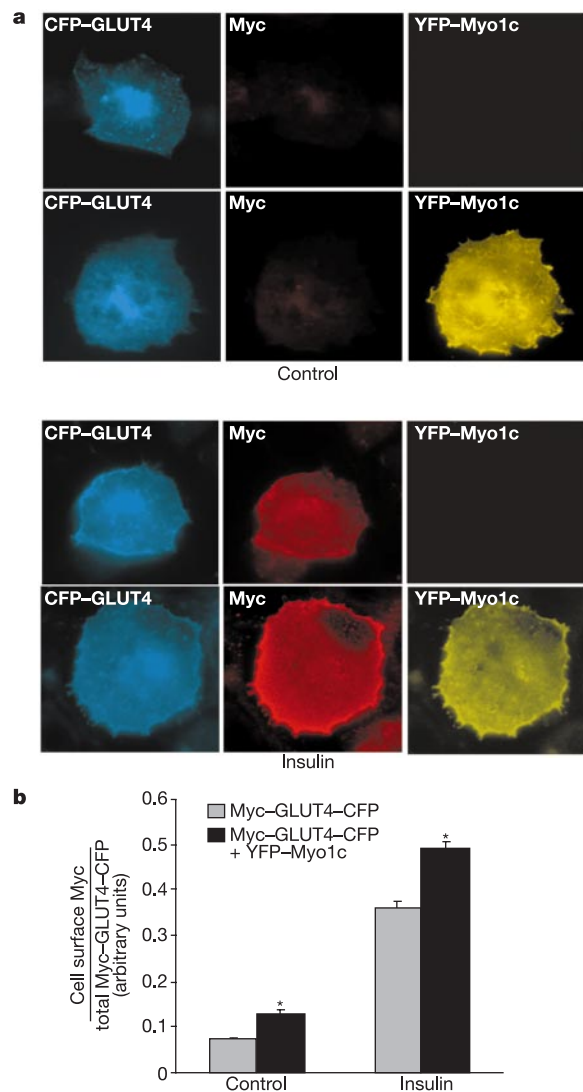


Figure 5 Expression of Myo1c-YFP in differentiated 3T3-L1 adipocytes potentiates insulin-stimulated Myc-GLUT4-CFP translocation. **a**, Differentiated 3T3-L1 adipocytes expressing Myc-GLUT4-CFP and YFP-Myo1c were stimulated with insulin. The cells were then fixed and stained with 9E10 antibodies against Myc. In both conditions, top images show a cell expressing only Myc-GLUT4-CFP and bottom images show a cell expressing both Myc-GLUT4-CFP and YFP-Myo1c. Data are representative of three independent experiments. **b**, Quantification of cell-surface anti-Myc signal (Alex 594) in the 3T3-L1 adipocytes. The arbitrary unit represents the ratio of the cell-surface Myc signal to the total CFP signal in each cell. Asterisk, $P < 0.0001$.

Methods

Plasmids and antibodies

The pCMV5–3XHAMyolc(T) plasmid was constructed by subcloning Myolc coding sequence encompassing residues 767–1,028, amplified by polymerase chain reaction (PCR), into the *KpnI* and *BamHI* restriction sites of pCMV5–3XHA in frame with the HA epitope. The plasmid expressing HAMyolb(T) was constructed similarly by amplifying the region encoding residues 747–1,041 in the Myolb complementary DNA. The plasmid encoding YFP–Myolc was constructed by cloning the PCR-amplified cDNA encoding Myolc into the *HindIII/BamHI* sites of the vector pEYFPC1 in frame with YFP. The Myc–GLUT4–EGFP plasmid has been described⁸. All of the constructs were sequenced and their expression was verified in COS-1 cells before they were transfected into 3T3-L1 adipocytes.

The rabbit polyclonal antibody against Myolc (Tu49) was a gift from M. Bahler (Munich). We also generated our own polyclonal antibodies against Myolc (henceforth referred to as anti-Myolc) using the protocol described²³. Antibodies against goat GLUT4 (C–20) were purchased from Santa Cruz Biotechnology and antibodies against myosin 2b were from Babco. The antibody against EHD2 was generated in rabbit using a glutathione S-transferase fusion protein of EHD2 as an antigen.

Cell culture, cell treatments and transfection

The 3T3-L1 fibroblasts were grown to confluency and differentiated as described²⁸. They were serum-starved in DMEM medium plus 0.5% bovine serum albumin for 4 h. After serum starvation, cells were either left untreated or treated with 5 μ M latrunculin B for 1 h, 30 μ M nocodazole for 1 h, or 100 nM wortmannin for 15 min before being treated with 100 nM insulin for 15 min as indicated in the figure legends.

Differentiated 3T3-L1 adipocytes were transfected by electroporation as described²⁸. The cells were then replated and allowed to recover for 24–36 h before serum starvation and stimulation with insulin. For cotransfections, 100 μ g of pCMV5–3XHAMyolc(T) or pCMV5–3XHAMyolb(T) and 50 μ g of Myc–GLUT4–EGFP were used. For expression of Myolc, 50 μ g of YFP–Myolc was electroporated in combination with 50 μ g of the plasmid encoding Myc–GLUT4–CFP.

Myc–GLUT4–EGFP translocation assays

For the Myc–GLUT4–GFP translocation assay, we transfected 3T3-L1 adipocytes with Myc–GLUT4–EGFP either alone or with HA–Myolc(T) or HA–Myolb(T). The cells were then treated and analysed as described⁸. The cell-surface anti-Myc signal was quantified by subtracting the nonspecific Alexa 594 signal in the area inside the cell from the total Alexa 594 signal.

Northern blotting

Total RNA from 3T3-L1 fibroblasts, differentiated adipocytes and different mouse tissues was isolated as described²⁸. We resolved 10 μ g of total RNA on a 1.2% denaturing gels and transferred it to a Nytran membrane as described²⁶. DNA probes corresponding to the cargo domain of Myolc (residues 767–1,028) and Myolb (residues 747–1,041) were prepared by PCR amplification, labelled with [α -³²P]dCTP and hybridized to the membrane at 42 °C. The phosphor screen was done with a Molecular Dynamics Storm 860 scanner.

Digital image analysis

Fluorescence microscopy was carried out with a IX170 inverted microscope (Olympus America) with a CCD (charge coupled device) camera (Roper Scientific) and Metamorph image processing software (Universal Imaging).

Whole-mount electron microscopy

Cells were seeded onto freshly prepared Formvar-coated gold grids and washed twice with wash buffer (0.1 M phosphate buffer, pH 7.2). The cells were then permeabilized in wash buffer plus 0.5 mg ml^{−1} saponin for 20 min at 4 °C. We washed the cells and fixed them in a mixture of 2% paraformaldehyde and 0.2% glutaraldehyde for 10 min at 4 °C, quenched them in wash buffer plus 50 mM NH₄Cl and washed them three times in fresh buffer.

Double immunogold labelling with different primary antibodies (anti-Myolc with C-20 or rabbit IgG with goat IgG; all antibodies used at 10 μ g ml^{−1}) and specific secondary antibodies conjugated to 6-nm and 12-nm gold particles was carried out on paraffin sheets at 25 °C. After immunogold labelling, the cell preparations were fixed with 2.5% glutaraldehyde in phosphate buffer (pH 7.2), rinsed three times with water, dehydrated through a graded series of ethanol to two changes of 100% ethanol, and dried to critical point in a CO₂ CPD (critical point drying) apparatus. Whole-mount grids were examined on a Philips CM 10 transmission electron microscope at 80 K.

siRNA-induced degradation of Myolc

The siRNAs purchased from Dharmacon were designed to target the following cDNA sequences: scrambled, 5'-CAGTCGCGTTTGGCAGTGG-3'; siRNA 2, 5'-AAGGCGTTGTACAGCCGACATT-3'; siRNA 6, 5'-AAGCTTCCAGACAGGGATCCATG-3'. Either 20 nmol of scrambled or siRNA 6 alone, or 10 nmol each of siRNA 2 and siRNA 6 were electroporated into differentiated 3T3-L1 adipocytes (for detailed methods see ref. 30). After electroporation the cells were reseeded in 24-well plates and allowed to rest for 24 h. Glucose uptake was measured as described²⁸. The fold increase was calculated after normalizing the basal uptake in scrambled siRNA-transfected cells to 1 and using that as a control. We used a paired *t*-test to analyse the percentage decrease in 2-deoxyglucose uptake in presence of siRNA 6 alone versus siRNA 6 and siRNA 2 combined. A portion of these cells was analysed for Myolc, myosin 2b and EHD2 by western blotting.

Received 2 August; accepted 10 October 2002; doi:10.1038/nature01246.

1. Czech, M. P. & Corvera, S. Signaling mechanisms that regulate glucose transport. *J. Biol. Chem.* **274**, 1865–1868 (1999).
2. Saltiel, A. R. & Kahn, C. R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **414**, 799–806 (2001).
3. Baumann, C. A. *et al.* CAP defines a second signalling pathway required for insulin-stimulated glucose transport. *Nature* **407**, 202–207 (2000).
4. Chiang, S. H. *et al.* Insulin-stimulated GLUT4 translocation requires the CAP-dependent activation of TC10. *Nature* **410**, 944–948 (2001).
5. Czech, M. P. Lipid rafts and insulin action. *Nature* **407**, 147–148 (2000).
6. Omata, W., Shibata, H., Li, L., Takata, K. & Kojima, I. Actin filaments play a critical role in insulin-induced exocytotic recruitment but not in endocytosis of GLUT4 in isolated rat adipocytes. *Biochem. J.* **346**, 321–328 (2000).
7. Wang, Q., Bilan, P. J., Tsakiridis, T., Hinek, A. & Klip, A. Actin filaments participate in the relocalization of phosphatidylinositol 3-kinase to glucose transporter-containing compartments and in the stimulation of glucose uptake in 3T3-L1 adipocytes. *Biochem. J.* **331**, 917–928 (1998).
8. Jiang, Z. Y., Chawla, A., Bose, A., Way, M. & Czech, M. P. A phosphatidylinositol 3-kinase-independent insulin signaling pathway to N-WASP/Arp2/3/F-actin required for GLUT4 glucose transporter recycling. *J. Biol. Chem.* **277**, 509–515 (2002).
9. Patki, V. *et al.* Insulin action on GLUT4 traffic visualized in single 3T3-L1 adipocytes by using ultrafast microscopy. *Mol. Biol. Cell* **12**, 129–141 (2001).
10. Guilherme, A. *et al.* Perinuclear localization and insulin responsiveness of GLUT4 requires cytoskeletal integrity in 3T3-L1 adipocytes. *J. Biol. Chem.* **275**, 38151–38159 (2000).
11. Lin, B. Z., Pilch, P. F. & Kandrak, K. V. Sortilin is a major protein component of GLUT4-containing vesicles. *J. Biol. Chem.* **272**, 24145–24147 (1997).
12. Laurie, S. M., Cain, C. C., Lienhard, G. E. & Castle, J. D. The glucose transporter Glut4 and secretory carrier membrane proteins (SCAMPs) colocalize in rat adipocytes and partially segregate during insulin stimulation. *J. Biol. Chem.* **268**, 19110–19117 (1993).
13. Kandrak, K. V. & Pilch, P. F. gp160, a tissue-specific marker for insulin-activated glucose transport. *Proc. Natl Acad. Sci. USA* **91**, 8017–8021 (1994).
14. Cain, C. C., Trimble, W. S. & Lienhard, G. E. Members of the VAMP family of synaptic vesicle proteins are components of glucose transporter-containing vesicles from rat adipocytes. *J. Biol. Chem.* **267**, 11681–11684 (1992).
15. Gillespie, P. G. *et al.* Myosin-I nomenclature. *J. Cell Biol.* **155**, 703–704 (2001).
16. Reizes, O., Barylko, B., Li, C., Sudhof, T. C. & Albanesi, J. P. Domain structure of a mammalian myosin I β . *Proc. Natl Acad. Sci. USA* **91**, 6349–6353 (1994).
17. Barylko, B., Wagner, M. C., Reizes, O. & Albanesi, J. P. Purification and characterization of a mammalian myosin I. *Proc. Natl Acad. Sci. USA* **89**, 490–494 (1992).
18. Wu, X., Jung, G. & Hammer, J. A. Functions of unconventional myosins. *Curr. Opin. Cell Biol.* **12**, 42–51 (2000).
19. Neuhäus, E. M. & Soldati, T. A myosin 1 is involved in membrane recycling from early endosomes. *J. Cell Biol.* **150**, 1013–1026 (2000).
20. Montes de Oca, G., Lézama, R. A., Mondragon, R., Castillo, A. M. & Meza, I. Myosin I interactions with actin filaments and trans-Golgi derived vesicles in MDMK cell monolayers. *Arch. Med. Res.* **28**, 321–328 (1997).
21. Holt, J. R. *et al.* A chemical-genetic strategy implicates myosin-1c in adaptation by hair cells. *Cell* **108**, 371–381 (2002).
22. Raposo, G. *et al.* Association of myosin I α with endosomes and lysosomes in mammalian cells. *Mol. Biol. Cell* **10**, 1477–1494 (1999).
23. Ruppert, C. *et al.* Localization of the rat myosin I molecules myr 1 and myr 2 and in vivo targeting of their tail domains. *J. Cell Sci.* **108**, 3775–3786 (1995).
24. Okada, T., Kawano, Y., Sakakibara, T., Haze, O. & Ui, M. Essential role of phosphatidylinositol 3-kinase in insulin-induced glucose transport and antilipolysis in rat adipocytes. Studies with a selective inhibitor wortmannin. *J. Biol. Chem.* **269**, 3568–3573 (1994).
25. Emoto, M., Langille, S. E. & Czech, M. P. A role for kinesin in insulin-stimulated GLUT4 glucose transporter translocation in 3T3-L1 adipocytes. *J. Biol. Chem.* **276**, 10677–10682 (2001).
26. Wu, X., Bowers, B., Wei, Q., Kocher, B. & Hammer, J. A. Myosin V associates with melanosomes in mouse melanocytes: evidence that myosin V is an organelle motor. *J. Cell Sci.* **110**, 847–859 (1997).
27. Wu, X., Bowers, B., Rao, K., Wei, Q. & Hammer, J. A. Visualization of melanosome dynamics within wild-type and dilute melanocytes suggests a paradigm for myosin V function in vivo. *J. Cell Biol.* **143**, 1899–1918 (1998).
28. Bose, A. *et al.* G α i1 signaling through ARF6 regulates F-actin mobilization and GLUT4 glucose transporter translocation to the plasma membrane. *Mol. Cell Biol.* **21**, 5262–5275 (2001).
29. Elbashir, S. M. *et al.* Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* **411**, 494–498 (2001).
30. Jiang, Z. Y. *et al.* Insulin signalling through Akt/PKB analyzed by siRNA-mediated gene silencing. *Proc. Natl. Acad. Sci.* (submitted).

Supplementary Information accompanies the paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements We thank M. Bahler for the Tu49 antiserum; E. Coudrier for the Myolb construct; P. Furcinitti for help with digital microscopy; J. Leschyk and G. Witman, G. Hendricks and the Core Proteomics Facility and the Core Electron Microscopy Facility of the NIH Diabetes and Endocrinology Research Center at the University of Massachusetts Medical School.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.P.C. (e-mail: Michael.Czech@umassmed.edu).

Helper T cells regulate type-2 innate immunity *in vivo*

Kanade Shinkai, Markus Mohrs* & Richard M. Locksley

Howard Hughes Medical Institute, Departments of Medicine and Microbiology/Immunology, University of California San Francisco, San Francisco, California 94143-0654, USA

Type-2 immunity requires orchestration of innate and adaptive immune responses to protect mucosal sites from pathogens. Dysregulated type-2 responses result in allergy or asthma¹. T helper 2 (T_H2) cells elaborate cytokines, such as interleukin (IL)-4, IL-5, IL-9 and IL-13, which work with toxic mediators of innate immune cells to establish environments that are inhospitable to helminth or arthropod invaders². The importance of T_H2 cells in coordinating innate immune cells at sites of inflammation is not known. Here we show that polarized type-2 immune responses are initiated independently of adaptive immunity. In the absence of B and T cells, IL-4-expressing eosinophils were recruited to tissues of mice infected with the helminth *Nippostrongylus brasiliensis*, but eosinophils failed to degranulate. Reconstitution with CD4 T cells promoted accumulation of degranulated IL-4-expressing cells, but only if T cells were stimulated with cognate antigen. Degranulation correlated with tissue destruction, which was attenuated if eosinophils were depleted. Helper T cells confer antigen specificity on eosinophil cytotoxicity, but not cytokine responses, so defining a novel mechanism that focuses tissue injury at sites of immune challenge.

Many of the polarized cytokines produced by T_H1 and T_H2 cells are important in coordinating innate immune responses at sites of inflammation. However, recent evidence suggests that innate immune cells are sufficient to induce type-1 immunity through Toll-like receptor (TLR)-mediated, pathogen-driven activation^{3–7}. To determine whether the innate arm of type-2 immunity can be initiated in the absence of an adaptive immune compartment, we used *Rag-1*^{−/−} mice, which lack T and B cells, to examine the capacity of the helminth *Nippostrongylus brasiliensis* to activate IL-4 expression from innate immune cells. *N. brasiliensis* powerfully induces a type-2 immune response during larval migration through the lung². IL-4-secreting cells appeared in the lungs of *Rag-1*^{−/−} animals infected with *N. brasiliensis*, peaking at day 8 post-infection, consistent with the peak of IL-4 production in infected wild-type mice (Fig. 1a). Compared with wild-type mice, the response was one-quarter as abundant and failed to be maintained after 12 days, presumably reflecting the absence of a reinforcing adaptive immune response. No IFNγ-producing cells were detected in infected wild-type or *Rag-1*^{−/−} mice, consistent with the expected polarized type-2 response (data not shown). Thus, innate cells are capable of expressing a polarized type-2 cytokine response in the absence of an adaptive immune compartment.

To track cells involved in the type-2 response, we used IL-4 reporter mice, designated *4get* mice (IL-4 green fluorescent protein-enhanced transcript), which contain a knockin IL-4 gene modified to express a bicistronic message linking IL-4 via an internal ribosomal entry site (IRES) element with enhanced green fluorescent protein (eGFP)⁸. Cells from *4get* mice faithfully report IL-4 expression without the need for re-stimulation while leaving endogenous IL-4 intact. *4get* and *4get* × *Rag-1*^{−/−} mice were infected with *N. brasiliensis* and kinetic analysis of eGFP fluorescence was determined by flow cytometry. In wild-type mice, eGFP⁺ cells accumulated first in lungs, and later in mesenteric lymph nodes and spleen, reflecting the migratory path of the worms

as they travel from subcutaneous sites of inoculation through the lungs, up the trachea, down the oesophagus and finally to the intestines⁸. In *4get* × *Rag-1*^{−/−} mice, the magnitude of the innate cytokine response was attenuated in the absence of adaptive immunity. Analysis of the numbers of non-B, non-T (lineage-marker negative) eGFP⁺ cells per lung or spleen of age- and size-matched, infected *4get* or *4get* × *Rag-1*^{−/−} mice revealed a defect in the later phase of recruitment and/or activation of IL-4-expressing innate immune cells in the absence of B and T cells (Fig. 1b). Although quantitatively similar at early time points, the number of IL-4-expressing innate cells was reduced to one-sixth on day six and to one-fortieth on day twelve in lungs of *4get* × *Rag-1*^{−/−} mice; a similar defect was observed in the spleen. However, as assessed by comparable mean fluorescence intensity (MFI) of eGFP, innate IL-4-expressing cells in lungs of wild-type and *Rag-1*^{−/−} mice

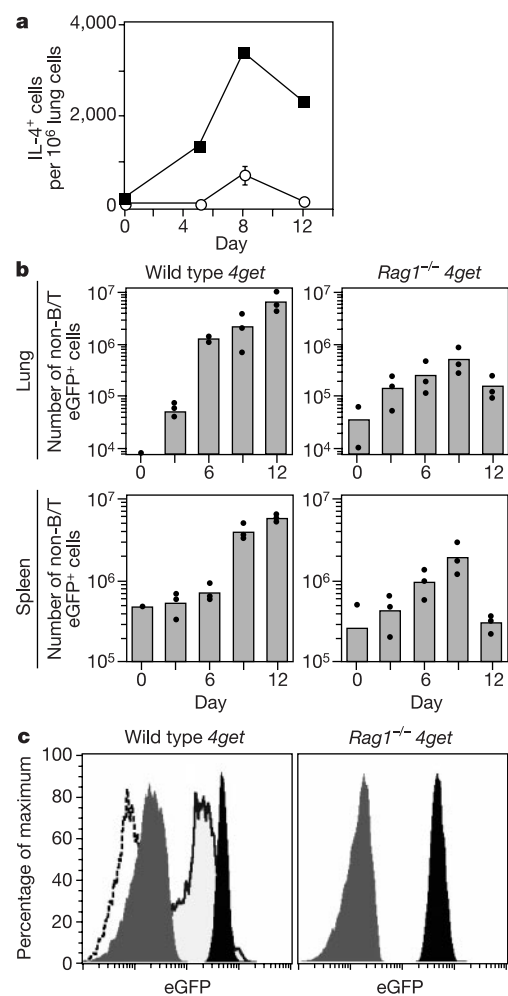


Figure 1 Activation of interleukin (IL-4)-expressing cells in the absence of adaptive immunity. **a**, Wild-type (filled squares) and *Rag-1*^{−/−} (open circles) mice were infected with *Nippostrongylus brasiliensis*. On designated days, lung cells were analysed for spontaneous IL-4 secretion by ELISPOT (enzyme-linked immunosorbent spot assay). **b**, *4get* and *4get* × *Rag-1*^{−/−} mice were infected with *N. brasiliensis*. On designated days, non-B, non-T cells from lung and spleen were analysed for spontaneous enhanced green fluorescent protein (eGFP) fluorescence by flow cytometry. Filled circles depict values in individual mice, bars depict group mean values. **c**, On day 9 post-infection, lung cells were analysed for mean fluorescence intensity of eGFP in CD4⁺ T cells (in *4get* mice) and non-B, non-T cells (in *4get* and *4get* × *Rag-1*^{−/−} mice). CD4⁺, eGFP⁺ (tinted with solid black line); CD4⁺, eGFP[−] (black dashed line); non-B/T, eGFP⁺ (solid black); non-B/T, eGFP[−] (solid grey). Data are representative of three comparable experiments.

* Present address: Trudeau Institute, Saranac Lake, New York 12983, USA.

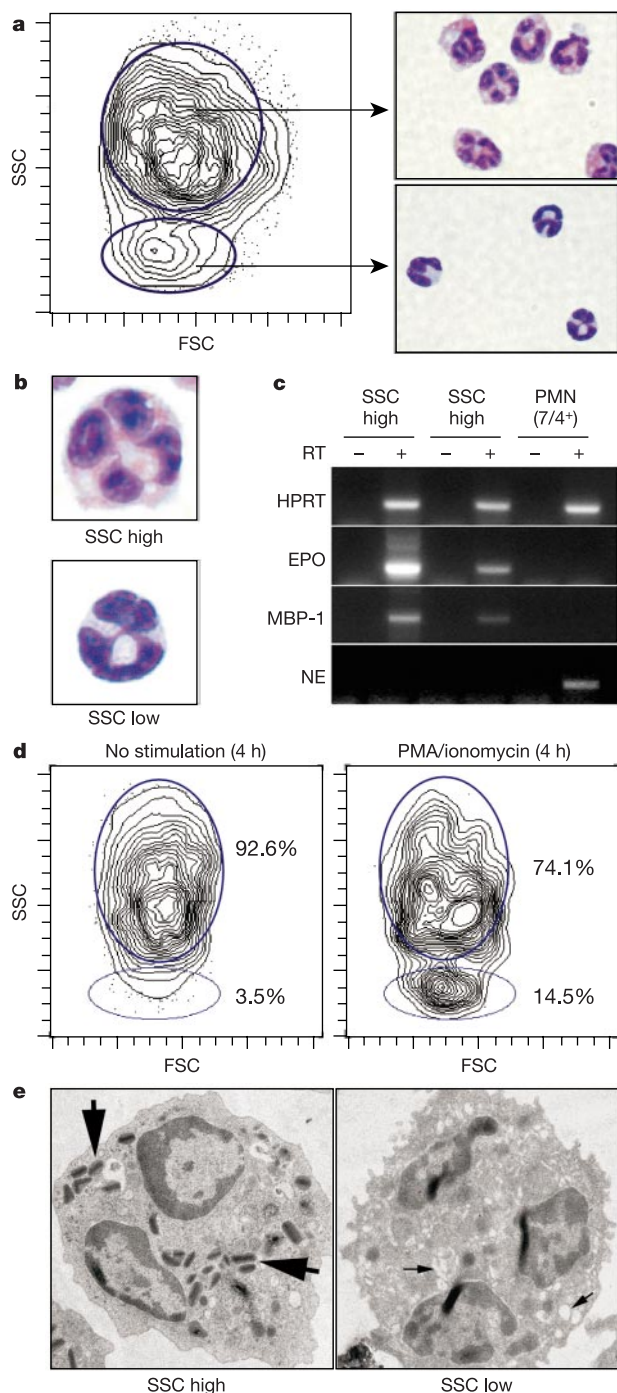


Figure 2 Subsets of eosinophils identified by flow cytometry. **a**, *4get* mice were infected with *N. brasiliensis*. eGFP⁺, non-B, non-T lung cells were analysed by flow cytometry (day 7 post-infection). Cells from indicated gates were isolated by cell sorting, centrifuged onto slides and stained with modified Wright–Giemsa (right panels). **b**, Detail of FSC (forward-scatter) cells sorted from SSC (side-scatter) high and low gates. **c**, eGFP⁺ SSC-high cells, eGFP⁺ SSC-low cells, and eGFP⁺ 7/4⁺ neutrophils (PMN) were sorted from lungs of infected *4get* mice (day 7 post-infection) and analysed by reverse transcriptase – polymerase chain reaction (RT–PCR) for expression of hypoxanthine–guanine phosphoribosyltransferase (HPRT), eosinophil peroxidase (EPO), major basic protein (MBP-1) and neutrophil elastase (NE). **d**, *4get* × *Rag-1*^{−/−} mice were infected with *N. brasiliensis*. On day 7, lung cells were maintained in media alone (left panel) or stimulated with 12-*O*-tetradecanoylphorbol-13-acetate (TPA) and ionomycin (right panel). After 4 h, eGFP⁺ cells were analysed by flow cytometry. **e**, Transmission electron microscopy of side-scatter high (left panel) and side-scatter low (right panel) non-B, non-T eGFP⁺ cells sorted from the lungs of infected *4get* mice (original magnification × 9168). Crystalloid-containing specific granules and cytoplasmic vacuoles indicated by thick and thin arrows, respectively. Data are representative of three comparable experiments.

demonstrated a robust cytokine-activating pathway that was independent of adaptive immunity (Fig. 1c).

Non-B, non-T eGFP⁺ cells were isolated from lungs of infected wild-type *4get* mice by flow cytometry for histological examination. Cells in the high side-scatter gate, which did not stain with the neutrophil-specific monoclonal antibody 7/4, were highly enriched for eosinophils (>90%), as characterized by lobulated nuclei and many dense, eosin-staining granules (Fig. 2a, b). The eGFP⁺ cells in the low side-scatter gate were also 7/4[−] and enriched for cells with morphologically similar nuclei, but contained only a perinuclear blush of lightly eosin-positive granules, and no dense granules. Both populations were viable, as determined by trypan blue and 7-AAD exclusion, and did not stain with annexin V (data not shown). Reverse-transcriptase polymerase chain reaction (RT–PCR) analysis of messenger RNA from each purified population confirmed comparable expression of eosinophil-specific genes, including major basic protein-1 (MBP-1) and eosinophil peroxidase (EPO), but not neutrophil-specific elastase, suggesting that both were eosinophils (Fig. 2c). Eosinophil-specific genes were not expressed by 7/4⁺ neutrophils. The hypothesis that the low side-scatter

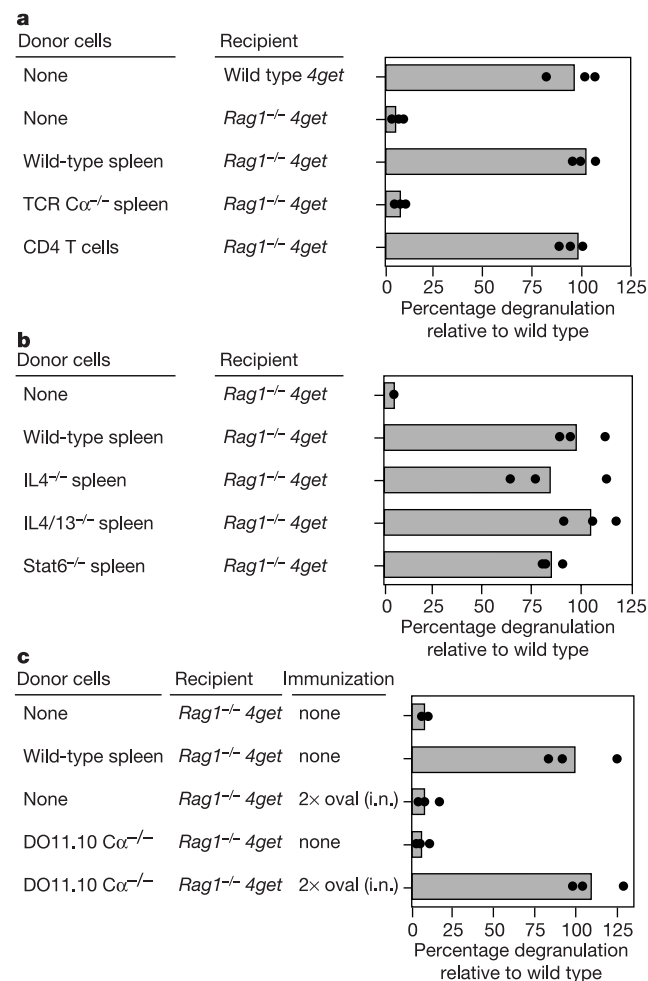


Figure 3 T helper cells regulate degranulation *in vivo*. **a**, Designated recipient mice were reconstituted with cells from the indicated donors and infected with *N. brasiliensis*. After 7 days, eGFP⁺, non-B, non-T lung cells were sorted onto slides, stained and analysed for degranulation by histological examination. Values are given relative to the percentage of cell degranulation in *4get* mice. Filled circles represent values in individual mice, bars represent group mean values. **b**, As in **a**, with source of donor cells as indicated. **c**, As in **a**, with source of donor cells as indicated. Where noted, cohorts of mice were treated additionally with intranasal (i.n.) ovalbumin twice before isolation of lung cells. Data are representative of three comparable experiments.

eosinophils might appear after degranulation of the granule-rich high side-scatter eosinophils was supported by the finding that treatment of the latter population with phorbol myristate acetate and ionomycin, agents that provoke eosinophil degranulation, resulted in the appearance of cells with the same side-scatter profile and morphological appearance as those in the low side-scatter gate (Fig. 2d, and data not shown). The appearance of degranulated eosinophils in the low side-scatter gate was concomitant with the reduction of eosinophils in the high side-scatter gate. Analysis by light microscope of eGFP⁺ cells (both low and high side-scatter gates) isolated from lungs of 4get mice infected with *N. brasiliensis* also revealed eosinophils at progressive stages of granule release (see Fig. A in Supplementary Information). Transmission electron microscopy of the eGFP⁺ high side-scatter cells isolated from lungs of infected 4get mice confirmed the typical lobulated nuclei and crystalloid core-containing specific granules of eosinophils (Fig. 2e). Ultrastructural study of the eGFP⁺ low side-scatter cells revealed intact cells with ruffled membranes containing similar lobulated nuclei but altered granules, including loss of the crystalloid core from specific granules and the accumulation of cytoplasmic vacuoles. Vesiculotubular organelles, a defining ultrastructural feature of degranulating human eosinophils⁹, were present in the cytoplasm of eGFP⁺, low side-scatter cells, as seen by high-power electron microscopy (see Fig. B in Supplementary Information).

In contrast to infected wild-type mice, in which up to 30% of the lung eGFP⁺ cells were degranulated by day 7, none (<1%) of the eGFP⁺ eosinophils in infected 4get × *Rag-1*^{-/-} mice were degranulated, even at periods as late as 14 days post-infection (Fig. 3a, and data not shown). Reconstitution of 4get × *Rag-1*^{-/-} mice with wild-type spleen cells at the time of infection was sufficient to restore the normal appearance of degranulated cells

in the lung by day 7 (Fig. 3a). Spleen cells from T-cell antigen receptor (TCR)-Cα^{-/-} mice did not restore degranulation, whereas purified wild-type CD4 T cells reconstituted the process, suggesting that CD4 T cells were sufficient to confer degranulation without the need for antibody. Although T_H2 cells presumably impart signals from CD4 T cells, spleen cells purified from IL-4-deficient, IL-4/IL-13-deficient and Stat6-deficient mice restored degranulation comparably to wild-type T cells (Fig. 3b). Thus, despite defects in T_H2 development and, in IL-4/IL-13- and Stat6-deficient mice, inability to expel *N. brasiliensis* after infection^{10,11}, T cells from these mice are competent to mediate the appearance of degranulated cells. Although DO11.10 CD4 T cells from ovalbumin-specific TCR transgenic mice on a TCR-Cα^{-/-} background were recruited to the lung, they were incapable of promoting degranulation unless mice were treated intranasally with ovalbumin (Fig. 3c, and data not shown). Thus, CD4 T cells are sufficient to reconstitute degranulation by a process that is antigen-dependent but antibody-independent.

Degranulation of eosinophils discharges toxic granule constituents, including MBP, EPO, eosinophil-derived neurotoxin, eosinophil cationic protein (ECP) and other inflammatory mediators¹². During infection with *N. brasiliensis*, the capacity of eosinophils to degranulate correlated with tissue injury. Lungs from infected wild-type mice were oedematous and visibly inflamed, in contrast to lungs from infected *Rag-1*^{-/-} mice (Fig. 4a, b). Reconstitution of *Rag-1*^{-/-} mice with CD4 T cells resulted in increased whole lung weight and inflammation comparable to wild-type mice. Histological examination revealed inflammation spilling into alveoli, epithelial injury and goblet cell mucus production in wild-type mice that was minimal in *Rag-1*^{-/-} mice, although peribronchial infiltrates were unaffected (Fig. 4c–f). Depletion of either helper T cells using anti-CD4 antibody or recruited eosinophils using IL-5-

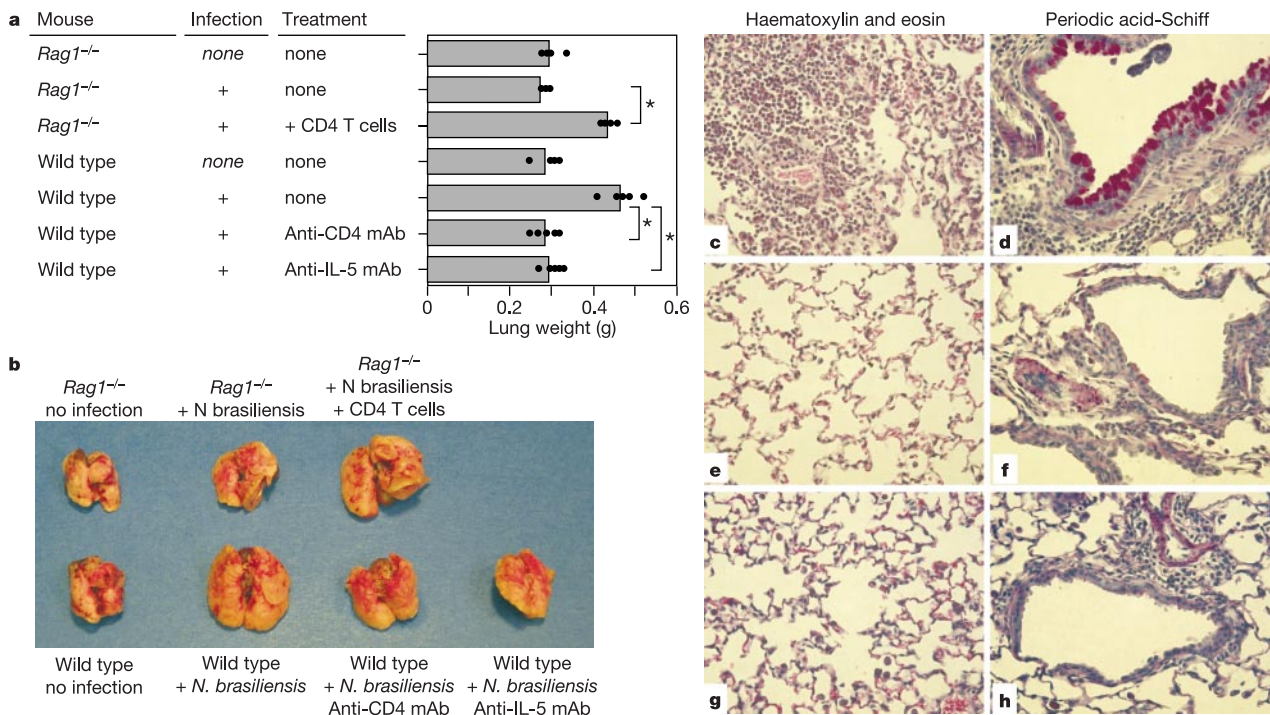


Figure 4 Tissue injury correlates with presence of both CD4 T cells and eosinophils. **a**, Designated mice were infected with *N. brasiliensis* after receiving adoptively transferred CD4 T cells or depleting antibodies to CD4 or IL-5. After 7 days, lungs were excised without perfusion and weighed. Filled circles depict values from individual mice, bars depict group mean values. Significant differences denoted by

asterisks ($P < 0.005$). **b**, Appearance of excised lungs from representative mice from each group. **c–h**, Haematoxylin and eosin (left) and periodic acid Schiff (right) stains of lung tissues on day 7 after infection in wild-type mice (**c**, **d**), *Rag-1*^{-/-} mice (**e**, **f**), and wild-type mice treated with anti-IL-5 antibody (**g**, **h**). Original magnification × 400. Data are representative of five mice per group.

neutralizing antibody resulted in reduced inflammation and attenuation of lung weight in infected mice (Fig. 4a, b, g, h, and data not shown).

These studies reveal interactions between innate and adaptive effector cells during the evolution of type-2 immune responses. First, eosinophils can activate IL-4 expression in the absence of adaptive immune cells *in vivo*. Thus, in agreement with previous observations in systems using *Schistosoma* egg and various environmental allergens, eosinophils express receptors capable of transducing signals from the environment to the IL-4 gene^{13,14}. Whether these receptors are pattern-recognition receptors that interact with specific pathogen moieties, or alternatively, that interact with ligands, cytokines, or chemokines induced in normal tissue in response to specific types of injury will require further study¹⁵. The preservation or default appearance of type-2 responses in MyD88-, Rip2- and IRAK4-deficient mice, however, indicates that these signals are distinct from those important in type-1 immunity⁴⁻⁷.

Eosinophil degranulation has been variably reported in models of allergic lung disease^{16,17}. In part, this may reflect histological uncertainty over the mechanism(s) underlying degranulation. Although cytolytic (necrotic) and classical exocytosis pathways have been described, our observations support a mechanism termed 'piecemeal' degranulation, in which dissolution of the crystalloid matrix core is followed by transport of specific granule constituents to the membrane for secretion, leaving empty cytoplasmic vacuoles¹⁸. Degranulated cells could be identified in our studies through the use of eGFP to label IL-4-expressing cells, but might otherwise have been overlooked. Studies in atopic patients support evidence for piecemeal degranulation of eosinophils after allergen exposure *in vivo*¹⁹. Unexpectedly, our quantitative approach revealed a profound role for antigen-activated CD4 T cells in mediating this process. CD4 T cells lacking components critical to T_H2 development, including IL-4, IL-13 and the signal-transducing element, Stat6, were competent to mediate degranulation. Previous evidence has suggested that IL-5 might mediate eosinophil degranulation and/or IL-4 expression^{16,20}, although these activities were completely dissociated in our experiments. IL-5 critically affects bone marrow generation of eosinophils and eosinophil survival, but transgenic mice over-expressing IL-5 do not demonstrate overt tissue cytotoxicity despite massive blood eosinophilia^{21,22}.

An important caveat to these studies remains the relationship of the low side-scatter cells to previously identified populations of IL-4-producing, non-T, non-B cells appearing in the spleen and lung of mice treated with anti-IgD or infected with *Nippostrongylus*²³. On the basis of morphological and phenotypic criteria, these cells were identified as basophils, which produced large amounts of IL-4 by an Fc receptor-mediated mechanism, reflecting antibody-dependent activation. Definitive identification of murine basophils remains difficult because CD117 (c-kit tyrosine kinase) expression is typically low or absent, cytoplasmic granules are relatively few and extensive vesicular networks have been described²⁴. Our interpretation that the low side-scatter gate cells are degranulated eosinophils is based on the homogeneous eGFP intensity in comparing high and low side-scatter cells, the precise matching of numbers of cells that moved from the high to the low gate after TPA/ionomycin treatment *in vitro*, the detection of eosinophil-specific gene transcripts in both populations, the antibody-independent expression of IL-4 by these cells in *Rag-1*^{-/-} mice, the light microscopic evidence for a spectrum of granulated populations, and the ultrastructural criteria that have been extrapolated from careful studies using human eosinophils. Further studies are required, however, to determine how CD4 T cells in wild-type mice function to recruit and/or activate IL-4-expressing basophils to tissues to participate in inflammatory type-2 responses.

CD4 T cell regulation of degranulation may represent an important checkpoint spatially restricting the release of tissue-destructive

inflammatory mediators contained in eosinophil granules. Eosinophil cationic proteins can generate transmembrane pores *in vitro*, and ECP and EPO catalyse formation of toxic intermediates from halides and oxygen, creating the potential for tissue cytotoxicity. Although these proteins are host protective in the setting of parasitic infection, the correlation between increased concentrations of MBP and EPO in bronchial lavage fluid and clinical exacerbations in asthma patients supports a cytotoxic role for eosinophil granular contents in disease¹². Although initial studies using neutralizing IL-5 antibody in asthma patients failed to demonstrate an impact of eosinophil depletion on bronchial hyper-reactivity²⁵, confounding effects including incomplete tissue eosinophil depletion and IL-5-independent pathways for eosinophil activation preclude definitive conclusions²⁶.

That effector T cells regulate the discharge of eosinophil granular contents, although unexpected, is conceptually similar to the role of IgE in mediating the discharge of mast cell contents. In both cell types, antigen-specific components of adaptive immunity regulate the ultimate decision mediating the discharge of toxic contents. The capacity to ameliorate asthma by targeting IgE²⁷ suggests that interventions abrogating the pathway by which antigen-specific T cells promote innate cell degranulation may prove fruitful therapeutic targets. Targeting helper T cells might attack dual components of the pathological response by both removing allergen-specific adaptive immunity and interfering with eosinophil cytotoxicity. □

Methods

Mice

Female BALB/c mice (Charles River Laboratories) were maintained in the University of California San Francisco specific pathogen-free facility and used at 6–8 weeks of age. Interleukin-4 GFP enhanced transcript (*4get*) mice⁸ were backcrossed three generations to BALB/c before use. Crossing N3 BALB/c *4get* mice onto N10 BALB/c *Rag-1*^{-/-} mice generated *4get* × *Rag-1*^{-/-} mice. *TCR- α* ^{-/-} mice were backcrossed ten generations onto BALB/c. N10 BALB/c *TCR- α* ^{-/-} mice were crossed to DO11.10 TCR transgenic mice²⁸. *IL-4*-, *IL-4/IL-13*^{-/-} and *Stat6*^{-/-} mice were on the BALB/c background.

Parasites and infection

Infective third-stage *N. brasiliensis* larvae were isolated from the faeces of infected rats by modified Baermann technique. After washing, 500 larvae in 0.2 ml saline were injected subcutaneously at the base of the tail. Infected mice were maintained on antibiotic-containing water.

Cell purification

On specified days, mice were killed and the lungs perfused with sterile phosphate buffered saline (PBS) through the right cardiac ventricle. Lungs and spleen were excised and mechanically dispersed into single-cell suspensions. Non-B, non-T cells were purified after labelling cells with antibodies to CD4 (YTS191.1, Caltag Laboratories), CD8 (5H10, Caltag), and CD19 (1D3, PharMingen) using flow cytometry (MoFlo Multi-laser Cytometer, Cytomation). Recovered cells were more than 98% CD4/CD8/CD19-negative.

Spleen and T cell reconstitutions and depletions

Eighty million spleen cells were injected intravenously in 0.2 ml PBS into *Rag-1*^{-/-} mice. Where indicated, CD4 T lymphocytes were enriched to 90–99% purity by either cell sorting or antibody- and complement-mediated lysis of CD8, class II major histocompatibility complex (MHC) and heat-stable antigen-bearing cells as described²⁹. Five million CD4 T cells were injected intravenously in 0.2 ml PBS into *Rag-1*^{-/-} mice. Cell depletions *in vivo* were performed by single intraperitoneal injection of 2 mg GK1.5 mAb (for CD4 T cells) or TRFK5 mAb (anti-IL-5, for eosinophils) concomitant with infection.

Cytokine analysis

Lung cells were washed and incubated at 10⁷ live cells per millilitre in culture medium (RPMI 1640 with 10% heat-inactivated foetal calf serum, 50 μ M 2-mercaptoethanol, 2 mM L-glutamine and 100 IU ml⁻¹ penicillin/100 μ g ml⁻¹ streptomycin) overnight at 37 °C in 5% CO₂ on anti-IL-4 (11B11)- or anti-IFN- γ (R46A2)-coated, 96-well ELISPOT (enzyme-linked immunosorbent spot assay) plates without re-stimulation. After 16 h, plates were washed and developed to detect the frequency of cytokine-expressing cells as described³⁰.

Histology

Cells purified by flow cytometry were centrifuged onto slides, fixed in methanol and stained using modified Wright-Giemsa. Eosinophil degranulation was assessed morphologically by enumerating at least 3,000 granulated or degranulated cells per individual mouse by light microscopy. For tissue histology, lungs were surgically excised

without cardiac perfusion, inflated and fixed in 10% formalin in PBS for 48 h, paraffin embedded and cut into 3-µm sections before staining with haematoxylin and eosin or periodic acid-Schiff stains. For ultrastructural microscopy, sorted cells were fixed in 2.5% glutaraldehyde, postfixed in 1% osmium tetroxide, dehydrated stepwise in ethanol and propylene oxide, and embedded in epon araldite. After sectioning, samples were mounted on grids and analysed on a Tecnai20 electron microscope. Four hundred cells per sample were analysed by transmission electron microscopy.

Reverse transcriptase – polymerase chain reaction analysis

Activated granulocytes were sorted from single-cell lung suspensions from infected mice (day 7 post-infection) based on forward- and side-scatter profiles and eGFP fluorescence. Sorted cells were >98% pure by histological examination. Total RNA was purified (Biotecx Laboratories), templated with oligo(dT) and random hexamer primers to make complementary DNA (Clontech Laboratories) and used for PCR analysis with the following primers: hypoxanthine–guanine phosphoribosyltransferase (HPRT): sense 5'-CCT GCT GGA TTA CAT CAA AGC ACT G - 3', anti-sense 5'-TCC AAC ACT TCG TGG GGT CCT-3'; mouse MBP-1: sense 5'-TCT ACT TCT GGC TCT TCT AGT CGGG-3', anti-sense 5'-GAC ACA GTG AGA TAG ACG CCA GTG; mouse EPO: sense 5'-ACT GTT TCC TGC TAG AGC TTT TGC-3', anti-sense 5'-AGA GTG CTG CTG TTC CTT CAG G-3'; mouse neutrophil elastase (NE): sense 5'-GGA ACT GAA CGT CAC GGT GGT C-3', anti-sense 5'-GTT TTG AAT CCA GTC CAC ATA C-3'.

Intranasal immunizations

Individual 4get × Rag-1^{-/-} recipients of wild-type, DO11.10 × TCR-α^{-/-}, or no spleen cells were immunized on days 3 and 6 post-infection. Chicken ovalbumin (Sigma Chemical Company) was administered to anaesthetized mice by the intranasal route using 2.5 mg each immunization delivered in 50 µl sterile PBS. PBS alone was used as a control.

In vitro degranulation of eosinophils

Lung cells were isolated as described above, washed and incubated at 10⁷ live cells/ml in culture medium at 37 °C in 5% CO₂ for 4 h in the presence or absence of phorbol 12-myristate 13-acetate (TPA, 40 pg ml⁻¹) and ionomycin (2 µg ml⁻¹) before analysis of forward- and side-scatter characteristics by flow cytometry.

Statistics

Analysis of lung weight differences was performed by matched pairs comparison (*t* test with four degrees of freedom).

Received 13 June; accepted 13 September 2002; doi:10.1038/nature01202.

- Wills-Karp, M. Immunologic basis of antigen-induced airway hyperresponsiveness. *Annu. Rev. Immunol.* **17**, 255–281 (1999).
- Finkelman, F. D. *et al.* Cytokine regulation of host defense against parasitic gastrointestinal nematodes: lessons from studies with rodent models. *Annu. Rev. Immunol.* **15**, 505–533 (1997).
- Akira, S. Toll-like receptors and innate immunity. *Adv. Immunol.* **78**, 1–56 (2001).
- Schnare, M. *et al.* Toll-like receptors control activation of adaptive immune responses. *Nature Immunol.* **29**, 47–50 (2001).
- Chin, A. I. *et al.* Involvement of receptor-interacting protein 2 in innate and adaptive immune responses. *Nature* **416**, 190–194 (2002).
- Kobayashi, K. *et al.* RICK/Rip2/CARDIAK mediates signalling for receptors of the innate and adaptive immune systems. *Nature* **416**, 194–199 (2002).
- Suzuki, N. *et al.* Severe impairment of interleukin-1 and Toll-like receptor signalling in mice lacking IRAK-4. *Nature* **416**, 750–754 (2002).
- Mohrs, M., Shinkai, K., Mohrs, K. & Locksley, R. M. Analysis of type 2 immunity in vivo with a bicistronic reporter. *Immunity* **15**, 303–311 (2001).
- Dvorak, A. & Weller, P. F. Ultrastructural analysis of human eosinophils. *Chem. Immunol.* **76**, 1–28 (2000).
- McKenzie, G. J., Fallon, P. G., Emson, C. L., Grecis, R. K. & McKenzie, A. N. J. Simultaneous disruption of interleukin (IL)-4 and IL-13 defines individual roles in T helper cell type 2-mediated responses. *J. Exp. Med.* **189**, 1565–1572 (1999).
- Urban, J. F. Jr *et al.* IL-13, IL-4Rα, and Stat6 are required for the expulsion of the gastrointestinal nematode parasite *Nippostrongylus brasiliensis*. *Immunity* **8**, 255–264 (1998).
- Walsh, G. M. Eosinophil granule proteins and their role in disease. *Curr. Opin. Hematol.* **8**, 28–33 (2001).
- Rumbley, C. A. *et al.* Activated eosinophils are the major source of Th2-associated cytokines in the schistosome granuloma. *J. Immunol.* **162**, 1003–1009 (1999).
- Justice, J. P. *et al.* Ragweed-induced expression of GATA-3, IL-4, and IL-5 by eosinophils in the lungs of allergic C57BL/6 mice. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **282**, L302–L309 (2002).
- Foster, P. S. *et al.* Elemental signals regulating eosinophil accumulation in the lung. *Immunol. Rev.* **179**, 173–181 (2001).
- Mould, A. W. *et al.* The effect of IL-5 and eotaxin expression in the lung on eosinophil trafficking and degranulation and the induction of bronchial hyperreactivity. *J. Immunol.* **164**, 2142–2150 (2000).
- Denzler, K. L. *et al.* Extensive eosinophil degranulation and peroxidase-mediated oxidation of airway proteins do not occur in a mouse ovalbumin-challenge model of pulmonary inflammation. *J. Immunol.* **167**, 1672–1682 (2001).
- Erjefält, J. S. & Persson, C. G. A. New aspects of degranulation and fates of airway mucosal eosinophils. *Am. J. Respir. Crit. Care Med.* **161**, 2074–2085 (2000).
- Karawajczyk, M. *et al.* Piecemeal degranulation of peripheral blood eosinophils. A study of allergic subjects during and out of pollen season. *Am. J. Respir. Cell Mol. Biol.* **23**, 521–529 (2000).
- Sabin, E. A., Kopf, M. A. & Pearce, E. J. *Schistosoma mansoni* egg-induced early IL-4 production is dependent upon IL-5 and eosinophils. *J. Exp. Med.* **184**, 1871–1878 (1996).
- Dent, L. A., Strath, M., Mellor, A. L. & Sanderson, C. J. Eosinophilia in transgenic mice expressing interleukin 5. *J. Exp. Med.* **172**, 1425–1431 (1990).
- Lee, N. A. *et al.* Expression of IL-5 in thymocytes/T cells leads to the development of a massive

eosinophilia, extramedullary eosinophilopoiesis, and unique histopathologies. *J. Immunol.* **158**, 1332–1344 (1997).

- Conrad, D. H., Ben-Sasson, S. Z., Le Gros, G., Finkelman, F. D. & Paul, W. E. Infection with *Nippostrongylus brasiliensis* or injection with anti-IgD antibodies markedly enhances Fc-receptor-mediated interleukin 4 production by non-B, non-T cells. *J. Exp. Med.* **171**, 1497–1508 (1990).
- Dvorak, A. M. Cell biology of the basophil. *Int. Rev. Cytol.* **180**, 87–236 (1998).
- Leckie, M. J. *et al.* Effects of an interleukin-5 blocking monoclonal antibody on eosinophils, airway hyper-responsiveness, and the late asthmatic response. *Lancet* **356**, 2144–2148 (2000).
- Foster, P. S. *et al.* Interleukin-5 and eosinophils as therapeutic targets for asthma. *Trends Mol. Med.* **8**, 162–167 (2002).
- Milgrom, H. *et al.* Treatment of allergic asthma with monoclonal anti-IgE antibody. rhuMAB-E25 study group. *N. Engl. J. Med.* **341**, 1966–1973 (1999).
- Murphy, K. M., Heimberger, A. B. & Loh, D. Y. Induction by antigen of interthymic apoptosis of CD4⁺ CD8⁺ TCR^{lo} thymocytes *in vivo*. *Science* **250**, 1720–1722 (1990).
- Fowell, D. J., Magram, J., Turck, C. W., Killeen, N. & Locksley, R. M. Impaired Th2 subset development in the absence of CD4. *Immunity* **6**, 559–569 (1997).
- Brown, D. R. *et al.* Beta-2-microglobulin-dependent NK1.1⁺ T cells are not essential for T helper cell 2 immune responses. *J. Exp. Med.* **184**, 1295–1304 (1996).

Supplementary Information accompanies this paper on Nature's website (<http://www.nature.com/nature>).

Acknowledgements The authors thank M.-L. Wong for assistance with electron microscopy, C. McArthur for expertise with cell sorting, San Francisco General Hospital Histopathology Laboratory Service for tissue histology, N. Flores, L. Stowring and Z.-E. Wang for technical assistance, B. Seymour and R. Coffman for parasites, R. Venkayya, D. Erle and A. McKenzie for mice, and J. Bluestone, J. Cyster, M. Kuhns, L. Lanier and laboratory members for critically reading the manuscript. This work was supported in part by funding from National Institute of Allergy and Infectious Diseases and National Heart, Lung and Blood Institute, National Institutes of Health. K.S. is supported by the UCSF Medical Scientist Training Program. R.M.L. is an Ellison Senior Scholar in Global Infectious Diseases.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to R.M.L. (e-mail: locksley@medicine.ucsf.edu).

Chloroplast avoidance movement reduces photodamage in plants

Masahiro Kasahara*, Takatoshi Kagawa*†, Kazusato Oikawa*‡, Noriyuki Suetsugu*‡, Mitsue Miyao§ & Masamitsu Wada*‡

* Division of Biological Regulation and Photobiology, National Institute for Basic Biology, Okazaki 444-8585, Japan

† Unit Process and Combined Circuit, PRESTO (Precursory Research for Embryonic Science and Technology), Japan Science and Technology Corporation, 1-8, Honcho 4-chome, Kawaguchi-city, Saitama 332-0012, Japan

‡ Department of Biological Sciences, Graduate School of Science, Tokyo Metropolitan University, Tokyo 192-0397, Japan

§ National Institute of Agrobiological Sciences, Tsukuba 305-8602, Japan

When plants are exposed to light levels higher than those required for photosynthesis, reactive oxygen species are generated in the chloroplasts and cause photodamage. This can occur even under natural growth conditions. To mitigate photodamage, plants have developed several protective mechanisms^{1–3}. One is chloroplast avoidance movement^{4–6}, in which chloroplasts move from the cell surface to the side walls of cells under high light conditions, although experimental support is still awaited^{7,8}. Here, using different classes of mutant defective in chloroplast avoidance movement, we show that these mutants are more susceptible to damage in high light than wild-type plants. Damage of the photosynthetic apparatus and subsequent bleaching of leaf colour and necrosis occur faster under high light conditions in the mutants than in wild-type plants. We conclude that chloroplast avoidance movement actually decreases the amount of light absorption by chloroplasts, and might therefore be important to the survival of plants under natural growth conditions.

To assess the effect of light intensity on chloroplast position in plant cells, fluence-rate–response curves for chloroplast movement were measured as increases in red light transmittance through leaves⁹ (Fig. 1a). The transmittance through wild-type leaves reached a maximum at about $500 \mu\text{mol m}^{-2} \text{s}^{-1}$ (Fig. 1a) within 30 min of treatment (data not shown). At this point, the chloroplasts were distributed at the anticlinal cell walls (Fig. 1b), indicating that they had undergone avoidance movement. The constant transmittance level, even under higher fluence rates, indicates that maximum chloroplast avoidance had already occurred at $500 \mu\text{mol m}^{-2} \text{s}^{-1}$.

It has recently been shown that chloroplast relocation movement in plants is mediated by blue-light receptors, the phototropins^{10–12}. Light transmittance, used here as a measure of the chloroplast density in sections of leaves through which a light beam is passed, does not change in leaves of plants defective in the *phototropin2* gene (*phot2-1*, formerly designated *cav1-1*), indicating that the chloroplast avoidance response does not occur in this mutant (Fig. 1a). Instead, chloroplasts in the *phot2-1* cells gather close to the cell surface under both weak and strong light conditions¹⁰ (Fig. 1b). In contrast, chloroplasts in leaf cells of the *phototropin1* (*phot1-5*, formerly designated *nph1-5*) mutant show an avoidance response

similar to that of wild-type plants (Fig. 1).

We have isolated an additional class of chloroplast movement mutant termed *chup1* (*chloroplast unusual positioning 1*) that contains an altered actin-binding protein. Because actin filaments have been shown to be involved in chloroplast movement^{13,14}, the CHUP1 protein might participate in this movement, rather than in the signal transduction pathway. *Chup1* mutant chloroplasts showed neither avoidance nor accumulation movements at any of the light intensities tested (Fig. 1a). The pattern of chloroplast distribution in *chup1-2* cells was quite different from that in *phot2-1* or wild-type cells. *Chup1-2* chloroplasts gathered at the bottom of cells, regardless of the light conditions, and even under low light conditions (Fig. 1b and Fig. 2b). The amount of light transmitted through *phot2-1* and *chup1-2* leaves remained relatively constant under increasing light intensity. Thus, chloroplasts in these mutants seem unable to decrease the amount of light exposure under strong light illumination by avoidance movement.

To determine whether *phot2-1* and *chup1-2* mutants are more susceptible to photodamage, mutant and wild-type plants grown under $100 \mu\text{mol m}^{-2} \text{s}^{-1}$ white light were exposed to continuous white light at $1,400 \mu\text{mol m}^{-2} \text{s}^{-1}$. Because the chloroplasts in these mutants are unable to undergo avoidance movement, this treatment might be expected to cause photobleaching, and this is what we observed. Leaves of *phot2-1* and *chup1-2* plants showed some signs of bleaching after 10 h of treatment but were severely bleached and necrotic after 22 h (Fig. 2a). In contrast, wild-type and *phot1-5* plants withstood this light stress over the 31-h duration of the treatment, although the old leaves became pale green and the young

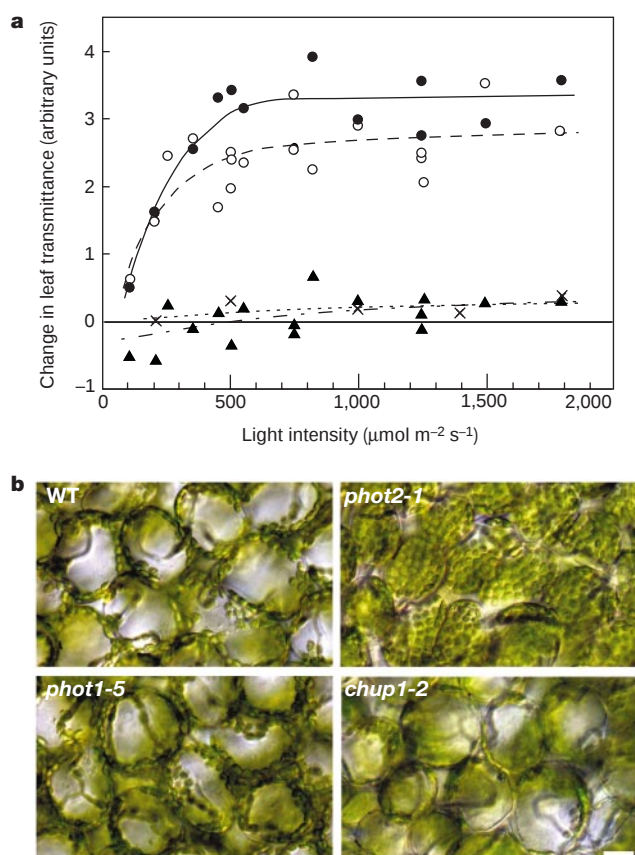


Figure 1 Chloroplast movement in *Arabidopsis* mutant cells. **a**, Changes in light transmittance through leaves of wild-type and mutant plants (4–5 weeks old) in response to various intensities of white light. Wild-type (filled circles), *phot2-1* (triangles), *phot1-5* (open circles) and *chup1-2* (crosses) mutant plants were irradiated with various intensities of white light for 1 h. The changes in leaf transmittance induced by light irradiation were plotted against light intensity. **b**, Distribution of chloroplasts in the mesophyll cells of wild-type and mutant leaves. Whole wild-type, *phot2-1*, *phot1-5* and *chup1-2* plants (4 weeks old) were irradiated with white light at $500 \mu\text{mol m}^{-2} \text{s}^{-1}$ for 1 h. The micrographs were taken from the adaxial side of the leaves. Scale bar, $20 \mu\text{m}$.

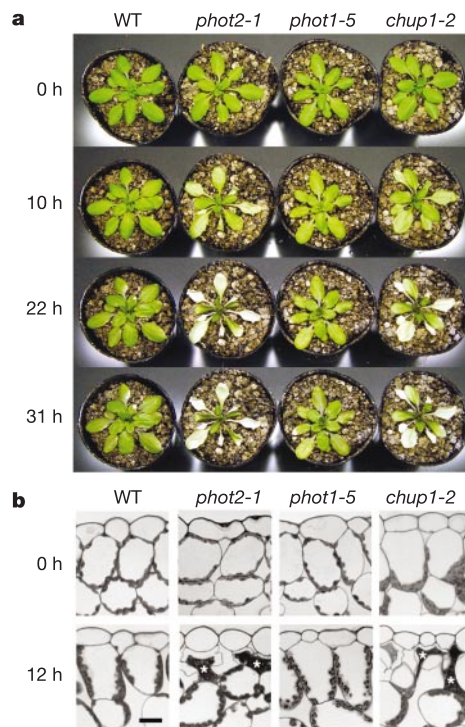


Figure 2 Phenotypes of plants exposed to continuous strong light. **a**, Wild-type (WT) and mutant plants exposed to continuous strong white light. Wild-type, *phot2-1*, *phot1-5* and *chup1-2* mutant plants (4 weeks old) were irradiated with white light at $1,400 \mu\text{mol m}^{-2} \text{s}^{-1}$ and were photographed before (0 h) and after irradiation for 10, 22 and 31 h. **b**, Cross-sections of wild-type and mutant leaves. Leaf sections were taken before (0 h) and after irradiation for 12 h with white light at $1,400 \mu\text{mol m}^{-2} \text{s}^{-1}$. Note the burst cells and evidence of chloroplast accumulation outside the burst cells (shown as asterisks). Scale bar, $20 \mu\text{m}$.

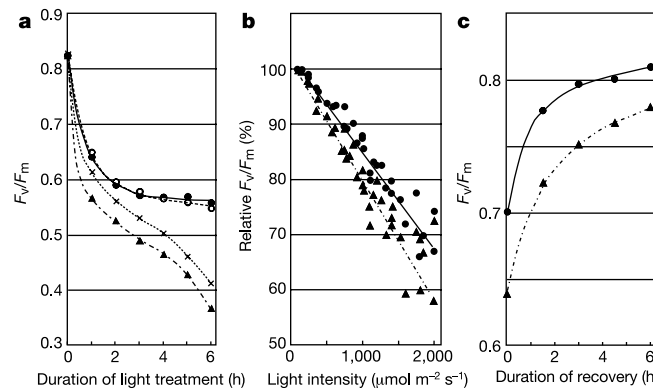


Figure 3 Inhibition and recovery kinetics of PSII photochemistry. **a**, Photoinhibition of PSII photochemistry by irradiation with strong light. Wild-type (filled circles), *phot2-1* (triangles), *phot1-5* (open circles) and *chup1-2* (crosses) mutant plants were exposed to white light at $1,400 \mu\text{mol m}^{-2} \text{s}^{-1}$ and F_v/F_m values were measured after the indicated durations of light exposure. Measurements were made from five or six different leaves on a single plant. Averages of data taken from three independent plants are shown. **b**, The sensitivity of PSII photochemistry to irradiation with various intensities of white light in wild-type and *phot2-1* leaves. The F_v/F_m values of leaves

of wild-type (filled circles) and *phot2-1* (triangles) plants were measured before and after irradiation for 1 h with various intensities ($100\text{--}2,000 \mu\text{mol m}^{-2} \text{s}^{-1}$) of white light. Each experimental point represents the mean value of five or six measurements from different leaves on the same plant. **c**, Recovery from photoinhibition of PSII in wild-type and *phot2-1* leaves. Wild-type and *phot2-1* mutant plants were treated for 1 h with high-intensity light ($1,400 \mu\text{mol m}^{-2} \text{s}^{-1}$) and left to recover under low light conditions ($70 \mu\text{mol m}^{-2} \text{s}^{-1}$).

leaves became dark green possibly because of high-light-induced anthocyanin accumulation (Fig. 2a). These wild-type and *phot1-5* leaves recovered their original colour several days after transfer back to low light conditions (data not shown). However, the photo-bleached leaves of *phot2-1* and *chup1-2* plants did not recover, and sections of these leaves revealed that some of the palisade cells near the adaxial leaf surface were damaged and had burst (Fig. 2b). Taken together, these results indicate that cells defective in chloroplast avoidance movement are indeed more susceptible than wild-type cells to photodamage.

To examine the susceptibility of the photosynthetic apparatus to photodamage, the chlorophyll fluorescence parameter F_v/F_m , the maximal quantum yield of photosystem II (PSII) photochemistry, was measured in wild-type and mutant plants exposed to high light (Fig. 3a). PSII is the primary target for photodamage^{15,16} and F_v/F_m can be a sensitive indicator of photoinhibition of photosynthesis¹⁷. After treatment with high light, F_v/F_m in wild-type plants declined rapidly to about 80% of the original value within 1 h, and then decreased gradually to reach 70% of the original value in 5 h (Fig. 3a). F_v/F_m for *phot1-5* plants showed a similar pattern to that of the wild type (Fig. 3a). In contrast to wild-type and *phot1-5* plants, F_v/F_m for *phot2-1* and *chup1-2* mutants continued to

decrease rapidly after a steep decline in the first hour to reach 45% and 50%, respectively, of the original value in 5 h (Fig. 3a). The extent of the decrease in F_v/F_m correlated well with the chloroplast area exposed to strong light illumination, seen as the green areas in Fig. 1b.

To study the relationship between F_v/F_m and light intensity, wild-type and *phot2-1* plants were given a 1-h treatment with white light of increasing intensity. F_v/F_m decreased linearly in wild-type samples with increasing light intensity, reaching ~70% of the original value (Fig. 3b). F_v/F_m for *phot2-1* plants also decreased linearly with light intensity, but was lower than that of the wild type at every fluence rate tested, with larger differences as the light intensity increased (Fig. 3b). These results indicate that chloroplast avoidance movement has a protective role against photoinhibition of photosynthesis even at relatively low light intensities. Recovery from photoinhibition under low light also differs between wild-type and *phot2-1* plants. After photoinhibition induced by strong light, F_v/F_m for wild-type plants largely recovered within 1 h, and then slowly recovered to reach the original level under weak light conditions (Fig. 3c). F_v/F_m for the *phot2-1* mutant also showed the same biphasic recovery pattern, but did not fully recover within 6 h, indicating that PSII in *phot2-1* was more severely damaged than

Table 1 Enzyme activities and antioxidant levels in wild-type and mutant plants before and after high light treatment

	Before high light treatment				
	CAT	APX (U per mg protein)	SOD	AsA ($\mu\text{mol per mg chlorophyll}$)	GSH
Wild type	206 $\pm$ 8	4.8 $\pm$ 0.1	29 $\pm$ 1	3.1 $\pm$ 0.4	0.14 $\pm$ 0.03
<i>phot2-1</i>	224 $\pm$ 17	4.8 $\pm$ 0.4	29 $\pm$ 4	3.1 $\pm$ 0.5	0.14 $\pm$ 0.03
<i>phot1-5</i>	206 $\pm$ 19	4.3 $\pm$ 0.3	31 $\pm$ 2	3.0 $\pm$ 0.7	0.15 $\pm$ 0.02
<i>chup1-2</i>	214 $\pm$ 1	4.5 $\pm$ 0.4	28 $\pm$ 1	3.0 $\pm$ 0.5	0.15 $\pm$ 0.03
	After high light treatment				
	CAT	APX (U per mg protein)	SOD	AsA ($\mu\text{mol per mg chlorophyll}$)	GSH
Wild type	192 $\pm$ 23	5.2 $\pm$ 0.3	32 $\pm$ 2	3.2 $\pm$ 0.5	0.15 $\pm$ 0.03
<i>phot2-1</i>	188 $\pm$ 12	5.2 $\pm$ 0.4	31 $\pm$ 1	3.2 $\pm$ 0.4	0.14 $\pm$ 0.02
<i>phot1-5</i>	173 $\pm$ 13	5.2 $\pm$ 0.7	30 $\pm$ 1	3.1 $\pm$ 0.6	0.16 $\pm$ 0.03
<i>chup1-2</i>	194 $\pm$ 18	5.3 $\pm$ 0.6	29 $\pm$ 4	3.0 $\pm$ 0.4	0.14 $\pm$ 0.04

Activities of catalase (CAT), ascorbate peroxidase (APX) and superoxide dismutase (SOD) and levels of ascorbate (AsA) and glutathione (GSH) were measured in wild-type and mutant leaves before and after high light treatment ($1,000 \mu\text{mol m}^{-2} \text{s}^{-1}$, 1 h) by the methods described previously¹⁸. All plants were grown under $100 \mu\text{mol m}^{-2} \text{s}^{-1}$ light before high light treatment. Data are mean $\pm$ s.d. ($n = 3$).

in wild-type plants, and took longer to recover completely.

Taken together, these results indicate that the photosynthetic apparatus of plants defective in the chloroplast avoidance movement are more susceptible to photoinhibition than in the wild type. The chlorophyll content and the chlorophyll fluorescence parameters, such as photochemical or non-photochemical quenching and the effective quantum yield ($\Delta F/F_m'$), were essentially the same in wild-type and mutant plants before high light treatment (data not shown). Therefore, the efficiency and functionality of the photosynthetic apparatus, including energy dissipation by non-photochemical quenching, seems to be unaffected by these mutations. The non-photochemical quenching values after mild light treatment (white light at $500 \mu\text{mol m}^{-2} \text{s}^{-1}$ for 1 h) of the mutant plants were comparable to or slightly higher than those of wild-type plants (data not shown), an indication that energy dissipation was upregulated by high light in similar ways in wild-type and mutant plants. Moreover, the reactive oxygen-scavenging capacity in these leaves, particularly the activities of scavenging enzymes (catalase, ascorbate peroxidase and superoxide dismutase) and the levels of antioxidants (ascorbate and glutathione)¹⁸, did not differ between wild-type and mutant plants (Table 1). Furthermore, the efficiency of gas exchange for photosynthesis under high light does not seem to differ either, because there is no difference in stomatal opening patterns between wild-type, *phot1-5* and *phot2-1* plants¹⁹.

We conclude that enhanced photoinhibition in the mutants can be attributed solely to the defect in chloroplast relocation in response to high light, rather than to defects in other photoprotective mechanisms such as energy dissipation and reactive-oxygen-scavenging systems or in photosynthesis itself. Chloroplast avoidance movement actually has a role in reducing light absorption by photosystems under high light. In the mutants, the enhanced photoinhibition of photosynthesis accelerates subsequent photo-damage, resulting in enhanced cell damage after prolonged exposure to high light (Fig. 2).

Since the discovery of chloroplast avoidance movement, its physiological importance as a mechanism for photoprotection has been unclear⁵. However, our findings show that photosynthesis is more susceptible to photodamage in the absence of chloroplast avoidance movement than with it. Further, exposure of plants to prolonged high light causes increasingly rapid and severe cell damage. The fact that chloroplast avoidance movement has been found in all the plants tested so far, including algae, mosses, ferns and seed plants⁴, supports the idea that this avoidance movement is important for plant survival under natural growth conditions. To avoid stress under high light, plants have developed multiple photoprotective mechanisms, including the adjustment of the size of light-harvesting antennae^{20,21}, non-photochemical quenching to dissipate absorbed light energy as heat², and scavenging systems for reactive oxygen species¹. We suggest that chloroplast avoidance movement should be added to the list of mechanisms with a crucial role in protecting plants from photodamage under high light. □

Methods

Plant materials

Arabidopsis seeds were sown on medium-saturated vermiculite in plastic pots 6 cm diameter²² and overlaid with one layer of fine vermiculite. After vernalization at 4 °C for 3 days, plants were cultivated at $100 \mu\text{mol m}^{-2} \text{s}^{-1}$ under a photoperiod of 16 h light and 8 h dark at 22 °C and watered every several days. Light was provided by three 40-W white fluorescent tubes (FL40SW; NEC, Tokyo, Japan) and two 40-W red-light-enriched fluorescent tubes (FL40SPG; National Co. Ltd, Japan). All light treatments were performed with a 3.6-kW xenon lamp (Ushio, Japan) in a growth chamber at 22 °C and 50% humidity. The light intensity was measured by a quantum sensor (LI-190SB; LI-COR, Lincoln, NE, USA). Plants 4–5 wk old were used for all experiments.

Leaf light transmittance

Leaf light transmittances were measured in at least four different areas of each leaf with equipment assembled from a light-emitting diode (GL5UR3K; Sharp Corp., Osaka, Japan), a photodiode (S1227-66BR; Hamamatsu Photonics K.K., Hamamatsu, Japan), a

voltmeter and a power supply. Although initial light transmittance was somewhat different among leaves (fluctuating from 4% to 8% of the measuring beam), the transmittance values before and after light treatments were roughly proportional (data not shown). The final values of transmittance plotted on the y-axis of Fig. 1a were therefore estimated as 5% of initial transmittance. One unit on the y-axis in Fig. 1a represents a 1% increase in leaf transmittance over the initial value of 5%.

Leaf sections

Rosette leaves of *Arabidopsis* irradiated with high-intensity white light ($1,400 \mu\text{mol m}^{-2} \text{s}^{-1}$) for 0 or 12 h were fixed with 1% glutaraldehyde in 0.1 M sodium phosphate buffer, pH 7.0, at 25 °C, and then with 4% OsO₄ on ice. The fixed tissues were dehydrated with acetone and embedded in Spurr resin. Thin sections (1 μm thick) were cut with a Reichert Ultramicrotome (Ultracut; Reichert-Jung, Vienna, Austria) and stained with toluidine blue.

Chlorophyll fluorescence measurements

Chlorophyll fluorescence was measured with a fluorometer (PAM-2000; Heinz Walz GmbH, Effeltrich, Germany). Before taking the F_v/F_m measurements, plants (4–5 wk old) were dark-adapted for 15 min (Fig. 3a and c) or 30 min (Fig. 3b). The values shown in Fig. 3b are those after light treatment relative to those before light treatment. The F_v/F_m values before light treatment were 0.81–0.83 in both wild-type and mutant plants.

Received 17 June; accepted 7 October 2002; doi:10.1038/nature01213.

- Asada, K. The water-water cycle in chloroplasts: scavenging of active oxygens and dissipation of excess photons. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 601–639 (1999).
- Niyogi, K. K. Photoprotection revisited: genetic and molecular approaches. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50**, 333–359 (1999).
- Demmig-Adams, B. & Adams, W. W. III Photoprotection and other responses of plants to high light stress. *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **43**, 599–626 (1992).
- Wada, M. & Kagawa, T. *Photomovement* (eds Häder, D.-P. & Lebert, M.) 897–924 (Elsevier, Amsterdam, 2001).
- Haupt, W. & Scheuerlein, R. Chloroplast movement. *Plant Cell Environ.* **13**, 595–614 (1990).
- Zurzycki, J. *The Blue Light Syndrome* (ed. Senger, H.) 50–68 (Springer, Berlin, 1980).
- Brugnoli, E. & Björkman, O. Chloroplast movements in leaves: influence on chlorophyll fluorescence and measurements of light-induced absorbance changes related to ΔpH and zeaxanthin formation. *Photosynth. Res.* **32**, 33–35 (1992).
- Park, Y.-I., Chow, W. S. & Anderson, J. M. Chloroplast movement in the shade plant *Tradescantia albiflora* helps protect photosystem II against light stress. *Plant Physiol.* **111**, 867–875 (1996).
- Trojan, A. & Gabrys, H. Chloroplast distribution in *Arabidopsis thaliana* (L.) depends on light conditions during growth. *Plant Physiol.* **111**, 419–425 (1996).
- Kagawa, T. et al. *Arabidopsis* NPL1: a phototropin homolog controlling the chloroplast high-light avoidance response. *Science* **291**, 2138–2141 (2001).
- Sakai, T. et al. *Arabidopsis* nph1 and npl1: blue light receptors that mediate both phototropism and chloroplast relocation. *Proc. Natl Acad. Sci. USA* **98**, 6969–6974 (2001).
- Jarillo, J. A. et al. Phototropin-related NPL1 controls chloroplast relocation induced by blue light. *Nature* **410**, 952–954 (2001).
- Kadota, A. & Wada, M. Photoorientation of chloroplasts in protonemal cells of the fern *Adiantum* as analyzed by use of a video-tracking system. *Bot. Mag. Tokyo* **105**, 265–279 (1992).
- Talaka, M. & Gabrys, H. Influence of calcium on blue-light-induced chloroplast movement in *Lemna trisulca* L. *Planta* **189**, 491–498 (1993).
- Powles, S. B. Photoinhibition of photosynthesis induced by visible light. *Annu. Rev. Plant Physiol.* **35**, 15–44 (1984).
- Barber, J. & Andersson, B. Too much of a good thing: light can be bad for photosynthesis. *Trends Biochem. Sci.* **17**, 61–66 (1992).
- Björkman, O. & Demmig, B. Photon yield of O₂ evolution and chlorophyll fluorescence characteristics at 77 K among vascular plants of diverse origins. *Planta* **170**, 489–504 (1987).
- Miyagawa, Y., Tamoi, M. & Shigeoka, S. Evaluation of the defense system in chloroplasts to photooxidative stress caused by paraquat using transgenic tobacco plants expressing catalase from *Escherichia coli*. *Plant Cell Physiol.* **41**, 311–320 (2000).
- Kinoshita, T. et al. *phot1* and *phot2* mediate blue light regulation of stomatal opening. *Nature* **414**, 656–660 (2001).
- Escoubas, J. M., Lomas, M., LaRoche, J. & Falkowski, P. G. Light intensity regulation of cab gene transcription is signaled by the redox state of the plastoquinone pool. *Proc. Natl Acad. Sci. USA* **92**, 10237–10241 (1995).
- Sukenik, A., Wyman, K. D., Bennett, J. & Falkowski, P. G. A novel mechanism for regulating the excitation of photosystem II in a green alga. *Nature* **327**, 704–707 (1987).
- Haughn, G. W. & Somerville, C. Sulfonylurea-resistant mutants of *Arabidopsis thaliana*. *Mol. Gen. Genet.* **204**, 430–434 (1986).

Acknowledgements We thank J. Silverthorne (University of California, Santa Cruz) and S. Christensen (Tokyo Metropolitan University) for critical reading of the manuscript, and N. Murata (National Institute for Basic Biology, Japan) for use of the PAM-2000 system. This work was partly supported by BRAIN (Program for Promotion of Basic Research Activities for Innovative Biosciences) and a Grant-in-Aid for Scientific Research from the Ministry of Education, Sports, Science and Technology (MEXT) of Japan to M.W., and by PRESTO, Japan Science and Technology Corporation, Japan, to T.K.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.W. (e-mail: wada-masamitsu@c.metro-u.ac.jp). The *chup1* sequence has been deposited in DDBJ/GenBank/EBI under accession number AB087408.

Role for a *Drosophila* Myb-containing protein complex in site-specific DNA replication

Eileen L. Beall^{*,†}, J. Robert Manak^{‡,†}, Sharleen Zhou^{*}, Maren Bell^{*}, Joseph S. Lipsick[‡] & Michael R. Botchan^{*}

^{*} Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA

[‡] Departments of Pathology and Genetics, Stanford University School of Medicine, Room L216, 300 Pasteur Drive, Stanford, California 94305-5324, USA

[†] These authors contributed equally to this work

There is considerable interest in the developmental, temporal and tissue-specific patterns of DNA replication in metazoans^{1,2}. Site-specific DNA replication at the chorion loci in *Drosophila* follicle cells leads to extensive gene amplification, and the organization of the *cis*-acting DNA elements that regulate this process may provide a model for how such regulation is achieved³. Two elements important for amplification of the third chromosome chorion gene cluster, *ACE3* and *Ori-β*, are directly bound by Orc^{4–6} (origin recognition complex), and two-dimensional gel analysis has revealed that the primary origin used is *Ori-β* (refs 7–9). Here we show that the *Drosophila* homologue of the Myb (Myeloblastosis) oncoprotein family is tightly associated with four additional proteins, and that the complex binds site-specifically to these regulatory DNA elements. *Drosophila* Myb is required *in trans* for gene amplification, showing that a Myb protein is directly involved in DNA replication. A *Drosophila* Myb binding site, as well as the binding site for another Myb complex member (p120), is necessary *in cis* for replication of reporter transgenes. Chromatin immunoprecipitation experiments localize both proteins to the chorion loci *in vivo*. These data provide evidence that specific protein complexes bound to replication enhancer elements work together with the general replication machinery for site-specific origin utilization during replication.

To identify proteins that bind to either *ACE3* or *Ori-β*, we fractionated *Drosophila* tissue culture nuclear extracts as outlined in Fig. 1. DNase I protection was used to assay site-specific *ACE3*- and *Ori-β*-binding proteins, and to follow their purification. The final glycerol gradient fractions contained five polypeptides (Fig. 1a) that co-eluted with binding activity for both DNAs (Fig. 1b, c) in multiple independent fractionation schemes from either Schneider L2 or Kc cell lines (data not shown). Utilizing peptide sequences from proteolysed purified protein, database searches identified *Drosophila* Myb (p85) and Caf1 p55 proteins, as well as three new *Drosophila* proteins (p40, p120, p130; Berkeley *Drosophila* Genome Project CG15119, CG6061, CG3480).

Drosophila Myb recognizes a highly conserved DNA sequence^{10,11}, but the specific binding properties of the glycerol gradient fractions might be more complex than that of Myb alone. We therefore tested whether any of the other proteins in our footprinting fractions might interact site-specifically with DNA. Each protein was produced individually and purified to homogeneity as either a (His)₆- or Flag-tagged protein using the baculovirus system. Only recombinant (r) Myb and rp120 bound to *ACE3* (Fig. 1d, lanes 2–6 and data not shown). rMyb protected nucleotides –471 to –511, and at higher concentrations protected –525 to –541, whereas rp120 protected –413 to –445, –525 to –541, and –575 to –603. However, the protection from the glycerol gradient fractions was more complex than the simple sums of the protections observed for these two purified proteins (Fig. 1d, compare lanes 3 and 6 to Fig. 1b, fractions 14–16). Moreover, rMyb on its own did not bind to *Ori-β*,

whereas p120 did (data not shown).

All five proteins co-immunoprecipitated together when any of the five antibodies were used, whereas a control antibody failed to immunoprecipitate any of the proteins (Fig. 2a). This association was not mediated through DNA, because ethidium bromide did not disrupt the interactions (Fig. 2b, compare lanes 2–4 with lanes 5–7). We obtained identical results using either 0–12-h embryo or ovary nuclear extracts (data not shown). Immunoblotting of fractions derived from the first two columns in our purification revealed that Myb co-fractionated only with the four other complex members (data not shown). No indication of free or other Myb forms was found. We therefore conclude that most of the Myb in these extracts is in a tight complex with the four additional proteins.

Antibodies against each of the five proteins individually super-shifted the nucleoprotein complex formed on *ACE3* DNA (Fig. 2c, compare lane 2 with lanes 3–7), whereas the control antibody did not (Fig. 2c, lane 8). Furthermore, when all of the recombinant proteins were co-expressed using baculovirus, the entire complex was isolated using nickel-chelate or Flag-tag affinity chromatography when only one subunit was tagged (data not shown).

As the *Drosophila* Myb complex bound to both *ACE3* and *Ori-β* *in vitro*, we tested whether the Myb complex directly interacted with Orc. Immunoprecipitation from Kc cell nuclear extracts showed that anti-Orc1 or anti-Orc2 antibodies co-immunoprecipitated

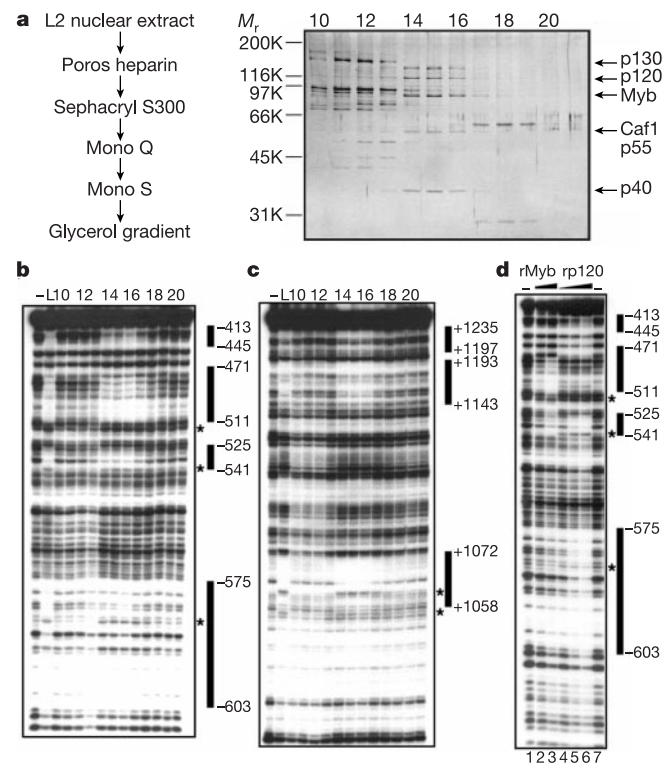


Figure 1 Purification of proteins binding to *ACE3* and *Ori-β*. **a**, Silver-stained gel showing the proteins from glycerol gradient fractions isolated using the scheme shown. The fraction numbers from the gradient (top), the five proteins that co-fractionated with the site-specific DNA-binding activity (right), and markers of relative molecular mass (M_r) (left) are indicated. **b**, DNase I protection by glycerol gradient fractions indicated on the top using a 5'-end-labelled *ACE3* DNA fragment. Minus sign indicates no protein; L shows glycerol gradient load from mono S peak. Black bars denote the regions of the probe that were protected from DNase I digestion; numbering on the right indicates the location of the sites within the 5'-untranslated region (5'-UTR) of the s18 chorion gene of *ACE3* (ref. 26). Asterisks denote DNase I hypersensitive sites. **c**, Same as **b** except that a 5'-end-labelled *Ori-β* fragment was used. **d**, Same as **b** except: lanes 1, 7, no protein; lanes 2–3, increasing amounts of recombinant (His)₆-Myb; lanes 4–6, increasing amounts of recombinant Flag-p120.

Orc1, 2 and 6 (Fig. 2d, lanes 1 and 2). Immunoprecipitations with anti-Myb antibodies, but not control IgG, co-immunoprecipitated Orc (Fig. 2d, lanes 3 and 4). Reciprocal experiments showed that anti-Orc2 antibodies co-immunoprecipitated the Myb complex (data not shown).

We performed chromatin immunoprecipitation assays on whole ovaries dissected from females that were aged to maximize the number of stage-10 egg chambers⁴ as a first step in exploring the role of the *Drosophila* Myb complex *in vivo*. We found that antibodies against Myb, Orc2 and p120 specifically precipitated *ACE3*-containing chromatin but did not precipitate the *actin* locus (Fig. 3a, lanes 3–5). Neither *ACE3* nor *actin* were precipitated with control IgG (Fig. 3a, lane 6). Thus, Myb, p120 and Orc are bound to *ACE3* in ovaries enriched for egg chambers undergoing chorion gene amplification. E2F-containing complexes⁵ can bind to Orc and are associated with *ACE3* in ovaries, but the location of this E2F *cis*-element is unknown. The interactions between the Myb complex, Orc and E2F proteins in sculpting the properties of the *ACE3* element will be critical to understanding how this element functions as a replication enhancer.

Small P element transgenes containing *ACE3* and *Ori-β* amplify efficiently at ectopic genomic sites only when both elements are

present⁹. We used a minimal replication reporter to assess the role of protein binding sites within *ACE3* to minimize the complications of redundant *cis*-elements. We used ‘suppressor of hairy wing binding sites’ (*SHWBS*) to insulate the transgenes from chromosomal position effects (Fig. 3b). Such reporters allow us to investigate, at various chromosomal positions, if the binding sites in *ACE3* for Myb and p120 are important *cis*-acting elements for amplification. We constructed transgenes that contained deletions of each of the binding sites identified by DNase I protection (Fig. 1b and Fig. 3b) and generated several transgenic lines for each deletion. Mutations abolished DNA binding of both recombinant Myb and the entire complex to the regulatory sequences in electrophoretic mobility shift assays (EMSAs) experiments (Supplementary Information). Deletion of either the Myb (–471 to –511) or one of the p120 binding sites (–413 to –445) resulted in severely reduced amplification in stage-13 egg chambers (Fig. 3c). Deletion of the other two p120 binding sites resulted in reduced levels of amplification, but not as severe as the sites shown in Fig. 3c (data not shown).

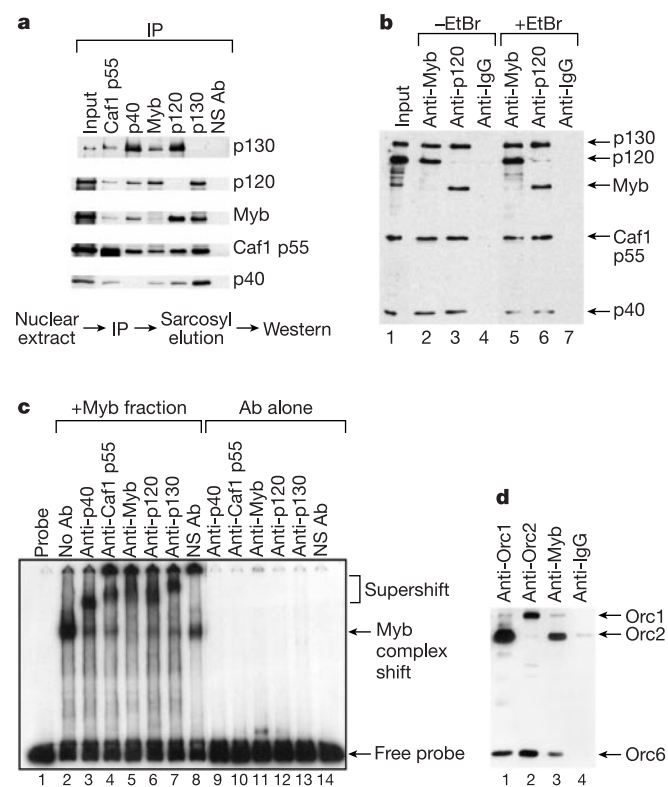


Figure 2 A Myb complex in *Drosophila*. **a**, Proteins from extracts were immunoprecipitated (IP) with a given antibody (top), the precipitates washed with RIPA buffer containing 300 mM KCl, and half of the 0.4% sarcosyl eluates were analysed using a single immunoblot successively probed with antibodies indicated (right): ‘input’ indicates 1/50 total extract. **b**, Same as in **a** except that the IPs were either performed with (lanes 5–7) or without (lanes 2–4) ethidium bromide (EtBr) in the extracts before precipitation. The immunoblot was performed with the mixture of antibodies. **c**, EMSA with mono S peak fraction and radiolabelled *ACE3* DNA (lane 2). Supershifts caused by given antibodies are indicated. NS, nonspecific control (affinity-purified anti-BPV-E2 polyclonal sera). Affinity-purified antibodies alone do not interact with DNA (lanes 9–13). **d**, IPs performed as in **a** in the presence of EtBr using the antibodies indicated in the top of the figure. Half of the sarcosyl eluate was simultaneously blotted for the presence of Orc1, 2 and 6 as indicated on the right. Lane 4, non-specific rabbit IgG.

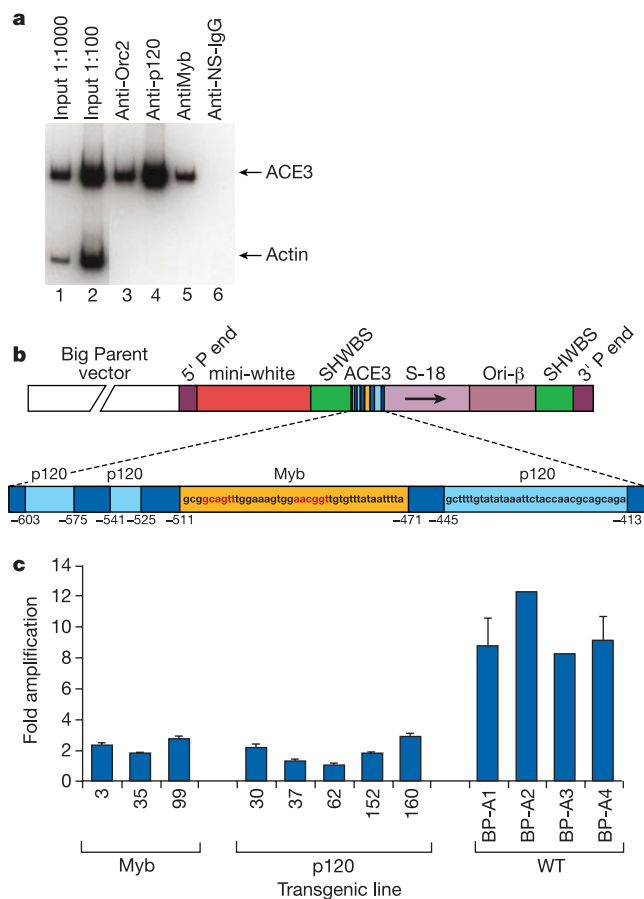


Figure 3 *Drosophila* Myb and p120 associate with *ACE3*, and binding sites are essential for amplification *in vivo*. **a**, *ACE3*- and *actin*-specific PCR following chromatin immunoprecipitation from whole ovaries using affinity-purified: anti-Orc2 (lane 3), anti-p120 (lane 4), anti-Myb (lane 5) or rabbit IgG (lane 6). PCR products from tenfold serial dilutions of the input DNA are shown in lanes 1 and 2. **b**, The P-element vector with insulator elements (*SHWBS*) surrounding the amplification unit. The *ACE3* region is expanded, showing the three p120 sites (pale blue) and the two Myb sites (yellow box with consensus sites in red) inferred from the DNase1 protection experiment. Three amplification reporters were used to create transgenic lines: wild type (WT), a reporter with base pairs –511 to –471 (Myb sites), or base pairs –445 to –413 (the most 3’ p120 site) deleted. **c**, Each blue bar (with strain no. given) represents data from an independent transgenic fly line; for each such line, at least two independent experiments were measured and each had side-by-side duplicates (mean $\pm$ s.d.).

The mature somatic follicle cells that surround the developing oocyte derive from a series of mitotic cell divisions followed by genome-wide endocycles³. At stage 10B, global DNA replication shuts down and the chorion loci on the X and third chromosomes (and two additional unidentified loci) begin locus-specific amplification. Amplification can be visualized by the incorporation of bromodeoxyuridine (BrdU) at four sub-nuclear foci using immunofluorescence microscopy, where the two largest foci represent incorporation on the X and third chromosomes¹².

As Orc2 also localizes to these foci, we stained ovaries with either anti-Myb or anti-Orc2 antibodies. We found that Myb is diffusely nuclear and not localized to the distinct sub-nuclear foci that contain Orc2 (compare Fig. 4a, lower left panel with Fig. 4c, upper right panel). Identical results were observed with the four other complex members (data not shown).

Drosophila Myb has been suggested to have a general role in S phase in several different tissues¹³. However, a direct role for

Drosophila Myb in replication at a specific location has not been demonstrated. To test the need for *Drosophila* Myb in *trans* for replication at the chorion loci, we used a directed mosaic system to generate green fluorescent protein (GFP)-negative, homozygous *Drosophila* Myb mutant clones in a heterozygous ovary^{14,15}. Figure 4a shows follicle cells from one such mutant clone in a stage-10 egg chamber that was stained with anti-Myb antibodies (lower left panel) and with propidium iodide for DNA (PI, upper right panel). There was negligible Myb protein expression in the GFP-negative clone (Fig. 4a, lower right panel).

In contrast to the surrounding normal cells, no subnuclear foci of BrdU incorporation were detected in *Drosophila* Myb mutant clones of stage-10 egg chambers (Fig. 4b). Double-labelling with anti-Orc2 antibodies and BrdU revealed that in *Myb* mutant cells, Orc2 was present at the sub-nuclear foci even though these cells lacked BrdU foci (Fig. 4c). Similar foci of Cdt1/Dup immunostaining were also present in *Myb* mutant cells (data not shown). We obtained

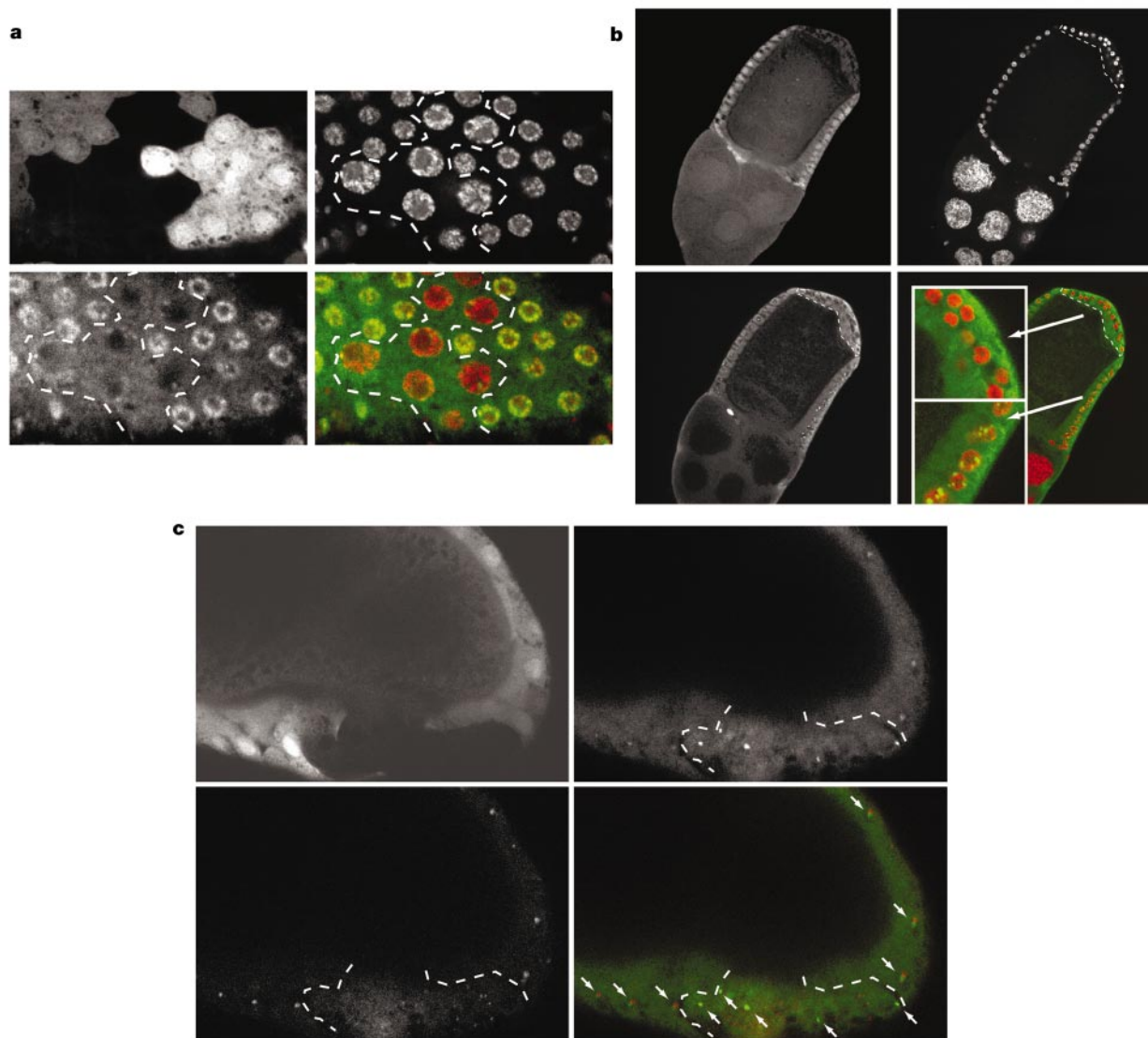


Figure 4 *Drosophila* Myb is required for chorion gene amplification. Stage-10 mosaic egg chambers from heterozygous *Myb* females (MH30 in **a** and **c**; MH107 in **b**). GFP-negative, *Myb* mutant mitotic clones (upper left in all three figures) are indicated with dotted lines. **a**, Propidium iodide (PI, upper right and red in lower right) and anti-Myb (lower left and green in lower right), showing an absence of Myb staining in the mutant clone. **b**, PI (upper right and red in lower right) and BrdU incorporation (lower left and green in lower right), showing the absence of subnuclear foci of BrdU incorporation in the *Myb* mutant clone (enlarged panels, lower right). **c**, Anti-Orc2 (upper right and green in lower right) and BrdU incorporation (lower left and red in lower right, downward-pointing arrows), showing normal Orc2 foci despite the absence of BrdU foci in the *Myb* mutant clone (upward-pointing arrows).

left and green in lower right), showing the absence of subnuclear foci of BrdU incorporation in the *Myb* mutant clone (enlarged panels, lower right). **c**, Anti-Orc2 (upper right and green in lower right) and BrdU incorporation (lower left and red in lower right, downward-pointing arrows), showing normal Orc2 foci despite the absence of BrdU foci in the *Myb* mutant clone (upward-pointing arrows).

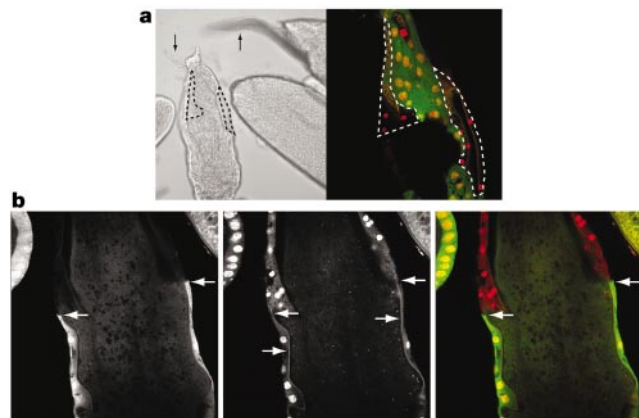


Figure 5 Absence of *Drosophila Myb* results in thin chorion and abnormal dorsal appendages. **a**, Left panel, differential interference contrast image of late-stage egg chambers from heterozygous *Myb* mutant, *AlkB*⁺ females. Arrow on left: abnormal dorsal appendage of the mosaic egg chamber; GFP-negative *Myb* mutant patches are outlined in black. Arrow on right: normal dorsal appendage of wild-type egg chamber. Right panel, mosaic egg chamber stained with PI (red) to visualize nuclei. The GFP (green)-negative *Myb* mutant patches are outlined in white. **b**, Egg chamber described in **a** showing GFP (left panel and green in right panel) and PI (centre panel and red in right panel). The chorion underlying the follicle cells in the GFP-positive, *Myb*-containing patch (right-pointing arrows) disappears at the border of the GFP-negative *Myb*-mutant patch (left-pointing arrows).

identical results with both of our previously described *Myb* mutants¹³: MH107 removes the promoter and 5'-untranslated region of *Myb* and the adjacent *AlkB* gene, and MH30 removes *AlkB*, *Myb* and a gene 3' of *Myb*. We also found that all of the MH107 phenotypes were complemented with a transgene that contained *Myb* without *AlkB*, but not with a transgene that contained *AlkB* without *Myb*¹³ (data not shown).

Together, these data demonstrate that in the absence of *Drosophila Myb*, members of the pre-replication-complex (RC) are present at the sub-nuclear foci but are unable to initiate detectable replication. One prediction from these results is that in late-stage mosaic egg chambers with *Myb* clones, patches of thin, fragile eggshell and thin dorsal appendages should result from reduced chorion gene expression. Figure 5a shows a mosaic egg chamber in which GFP-negative, *Myb* mutant patches in the regions that are normally responsible for dorsal appendage formation¹⁶ resulted in greatly reduced appendages (Supplementary Information). Note that as egg chambers age, the *Myb* mutant nuclei appear more compact. Using laser light scattering following propidium iodide staining, we found an absence of the underlying chorion beneath the follicle cells in the mutant patch (Fig. 5b, middle panel).

Studies of *Myb* family members have largely focused on their functions as transcription factors, though important targets for gene activation that might explain the role of these proteins in the cell cycle remain unclear. We show here that *Drosophila Myb* has a direct role in DNA replication. Our biochemical experiments imply that this protein functions in tight complex with four other proteins. Recent studies suggest that in a variety of tissues, but not all, *Drosophila Myb* seems to be important for S-phase progression^{13,17}. Our findings support a role for *Drosophila Myb* in tissue-specific and temporally defined DNA replication, much as enhancer proteins define site-specific transcription.

The genetic studies and, in particular, the mosaic analyses indicate that *Drosophila Myb* probably functions at a late step in the replication process, as *Orc2* and *Cdt1* were both localized at discrete foci within *Drosophila Myb* mutant stage-10B follicle cells. We infer that *Orc* and other general replication proteins were localized at *ACE3*. However, we do not know when in the develop-

mental process *Orc* appears at *ACE3* with regard to *Drosophila Myb* loss. Thus, *Drosophila Myb* perdurance after genomic deletion could certainly complicate any conclusions about the role of *Drosophila Myb* in establishing a pre-RC at the amplification foci. In any case, *Myb* family members interact with both acetylases and deacetylases¹⁸; thus, it is intriguing to consider the potential roles of this modification in either early or late steps of DNA replication initiation. □

Methods

Myb complex purification and recombinant protein expression

Schneider L2 or Kc cell nuclear extracts were prepared as described¹⁹. Details of the fractionation protocol will be provided upon request. Fractions containing the DNA-binding activity were loaded onto a 15–35% glycerol gradient, and peptides sequenced from the purified *Drosophila Myb* complex members²⁰ were used to identify the corresponding proteins from database searches. Complementary DNA clones were obtained from Research Genetics, sequenced, and following polymerase chain reaction (PCR) amplification, cloned into either pFastBacHTa (Gibco) to generate (His)₆-tagged proteins or pFastBac1 (Gibco) to generate untagged versions. Flag-tagged p120 was generated in pFastBac1 following PCR amplification of the 5' end with oligonucleotides encoding the Flag epitope. Recombinant proteins were purified from infected HI-5 nuclear extracts using standard methods.

Antibody purification and immunoprecipitations

PCR fragments for *Drosophila Myb*, p40 (CG15119), p120 (CG6061) and p130 (CG3480) were cloned into pVCH6²¹ and proteins purified from *Escherichia coli*. Polyclonal rabbit antibodies were affinity purified²² using cyanogen bromide (CNBR)-sepharose (Amersham) coupled to maltose binding protein (MBP)-*Myb*, (His)₆-p40, glutathione S-transferase (GST)-p120 or (His)₆-p130 and dialysed into PBS. Anti-Caf1 p55 antibodies were affinity purified from serum provided by J. Kadonaga. For IP and chromatin immunoprecipitation experiments, antibodies were isolated from sera using Gamma-bind plus beads (Amersham) and immunoprecipitations performed as described using affinity-purified rabbit IgG (Sigma) as a control²³. Where indicated, 50 µg ml⁻¹ ethidium bromide was included in the extracts before IP.

Biochemical assays

DNAse I protection experiments were performed as described²⁴, except that the probes were derived from the third chromosome chorion locus⁴. For EMSA, binding reactions were set up on ice and contained 0.1 M KCl, 0.2 mg ml⁻¹ bovine serum albumin (BSA), 5 mM MgCl₂, 2 µg ml⁻¹ poly [d(G-C)] [d(G-C)], and ~1 fmol of end-labelled *ACE3* probe. Reactions were incubated for 15 min at room temperature, affinity-purified antibodies were added, and further incubated for 15 min. Samples were electrophoresed for 5 h at 4.7 V cm⁻¹ in 4.2% (60:1 acrylamide:bisacrylamide) gels containing 0.5X TGE buffer (12.5 mM Tris, pH 8.3; 95 mM glycine; 0.5 mM EDTA), 0.02% ethyl phenyl-polyethylene glycol (NP-40; Shell), and 3% glycerol.

Chromatin immunoprecipitation

Chromatin immunoprecipitations were performed as described⁴ with the following modifications: whole ovaries were used instead of dissected stage-10 egg chambers, and immunoprecipitations were performed with 5 µl of Gamma-bind plus antibody beads. Fifteen cycles of PCR using *ACE3* and *actin* primers²⁵ and Southern blot hybridization with ³²P-labelled *ACE3* and *actin* probes were used to detect the amplified products.

Amplification assay

PCR products derived from oligonucleotides containing a deletion of the *Drosophila Myb* or p120 binding sites were used to replace the non-mutated *ACE3* fragment from the Big Parent vector⁹ and these constructs were transformed into *Drosophila*. The wild-type Big Parent lines were as described⁹. Amplification of the *ACE3* deletion transgenes was assayed⁹ with the following modification: the hybridization probe for the loading control was derived from a fragment of the *CG1115* single copy gene. For each independent transgenic line, at least four assays were performed, and the data presented as the mean ± s.d.

Genetics

The MH30 and MH107 alleles of *Drosophila Myb* were previously described¹³. Each allele was recombined with an X chromosome containing an FRT site at 18E (P[w⁺ mW.hs = FRT(whs)}9-2; Bloomington Stock Center). MH107, FRT18E was further recombined with an X chromosome P element containing a 9.8-kb *XbaI-BamHI Drosophila* genomic DNA fragment from 13F that contains the entire *G-beta* and *AlkB* genes, but only the 5' half of the *Drosophila Myb* gene¹³. Flies carrying y⁺, w¹¹¹⁸, P[w⁺ mC, Ubi-GFP.nls}, P[w⁺ mW.hs, FRT(whs)}9-2 or P[w⁺ mW.hs = en2.4- GAL4]e22c P[w⁺ mC = UAS-FLP1.D]}JD1 chromosomes were kindly provided by J. Duffy. Ovaries were dissected from adult female flies with genotypes: *Drosophila Myb* mutant, FRT18E/Ubi-GFP, FRT18E; e22c-GAL4, UAS-FLP. For testing the *Myb* mutation in the presence of wild-type *AlkB*, the *AlkB* + transgene was also present on the MH107, FRT18E chromosome. For testing the *AlkB* mutation in the presence of wild type *Myb*, we used homozygous MH107 females with a third chromosome P element containing a 6.6-kb *MluI-BamHI-MluI-BamHI* genomic DNA fragment containing all of *Myb* but only a small 3' fragment of *AlkB*¹³.

Antibody staining and BrdU labelling of ovaries

Ovaries were fixed and processed as described²⁵ except that all antibody incubations were performed in PBS with 0.1% Triton X-100 (PBT) and 1% BSA. BrdU labelling was performed as described¹² with the following modifications (kindly provided by B. Calvi): ovaries were washed 2 × 15 min in PBS + 0.6% Triton X-100, 2 × 15 min in DNase I buffer (66 mM Tris pH 7.5, 5 mM MgCl₂, 1 mM fresh 2-mercaptoethanol), then incubated in 25 units DNase I (Worthington)/0.5 ml DNase buffer at 37 °C for 30 min, washed 3 × 1 min then 1 × 30 min in PBT, blocked in PBT with 5% normal goat serum for at least 1 h, and stained with anti-BrdU monoclonal (Becton-Dickinson), affinity-purified anti-Myb polyclonal, and/or anti-Orc2 polyclonal⁴ antibodies followed by fluorophore-conjugated secondary antibodies (Molecular Probes). Ovaries were mounted in Vectashield with or without propidium iodide (Vector Laboratories) and visualized with a Nikon PCM confocal microscope.

Received 25 July; accepted 18 October 2002; doi:10.1038/nature01228.

1. Gilbert, D. M. Making sense of eukaryotic DNA replication origins. *Science* **294**, 96–100 (2001).
2. Bell, S. P. The origin recognition complex: from simple origins to complex functions. *Genes Dev.* **16**, 659–672 (2002).
3. Calvi, B. R. & Spradling, A. C. Chorion gene amplification in *Drosophila*: A model for metazoan origins of DNA replication and S-phase control. *Methods* **18**, 407–417 (1999).
4. Austin, R. J., Orr-Weaver, T. L. & Bell, S. P. *Drosophila* ORC specifically binds to ACE3, an origin of DNA replication control element. *Genes Dev.* **13**, 2639–2649 (1999).
5. Bosco, G., Du, W. & Orr-Weaver, T. L. DNA replication control through interaction of E2F-RB and the origin recognition complex. *Nature Cell Biol.* **3**, 289–295 (2001).
6. Chesnokov, I., Remus, D. & Botchan, M. Functional analysis of mutant and wild-type *Drosophila* origin recognition complex. *Proc. Natl Acad. Sci. USA* **98**, 11997–12002 (2001).
7. Heck, M. M. & Spradling, A. C. Multiple replication origins are used during *Drosophila* chorion gene amplification. *J. Cell Biol.* **110**, 903–914 (1990).
8. Delidakis, C. & Kafatos, F. C. Amplification enhancers and replication origins in the autosomal chorion gene cluster of *Drosophila*. *EMBO J.* **8**, 891–901 (1989).
9. Lu, L., Zhang, H. & Tower, J. Functionally distinct, sequence-specific replicator and origin elements are required for *Drosophila* chorion gene amplification. *Genes Dev.* **15**, 134–146 (2001).
10. Peters, C. W., Sippel, A. E., Vingron, M. & Klempner, K. H. *Drosophila* and vertebrate myb proteins share two conserved regions, one of which functions as a DNA-binding domain. *EMBO J.* **6**, 3085–3090 (1987).
11. Stober-Grasser, U. *et al.* The Myb DNA-binding domain is highly conserved in *Dictyostelium discoideum*. *Oncogene* **7**, 589–596 (1992).
12. Calvi, B. R., Lilly, M. A. & Spradling, A. C. Cell cycle control of chorion gene amplification. *Genes Dev.* **12**, 734–744 (1998).
13. Manak, J. R., Mitiku, N. & Lipsick, J. S. Mutation of the *Drosophila* homologue of the Myb protooncogene causes genomic instability. *Proc. Natl Acad. Sci. USA* **99**, 7438–7443 (2002).
14. Margolis, J. & Spradling, A. Identification and behavior of epithelial stem cells in the *Drosophila* ovary. *Development* **121**, 3797–3807 (1995).
15. Duffy, J. B., Harrison, D. A. & Perrimon, N. Identifying loci required for follicular patterning using directed mosaics. *Development* **125**, 2263–2271 (1998).
16. Suzanne, M., Perrimon, N. & Noselli, S. The *Drosophila* JNK pathway controls the morphogenesis of the egg dorsal appendages and micropyle. *Dev. Biol.* **237**, 282–294 (2001).
17. Fitzpatrick, C. A., Sharkov, N. V., Ramsay, G. & Katzen, A. L. *Drosophila* myb exerts opposing effects on S phase, promoting proliferation and suppressing endoreplication. *Development* **129**, 4497–4507 (2002).
18. Ganter, B. & Lipsick, J. S. Myb and oncogenesis. *Adv. Cancer Res.* **76**, 21–60 (1999).
19. Gossen, M., Pak, D. T., Hansen, S. K., Acharya, J. K. & Botchan, M. R. A *Drosophila* homolog of the yeast origin recognition complex. *Science* **270**, 1674–1677 (1995).
20. Ausubel, F. M., *et al.* *Current Protocols in Molecular Biology* (Wiley and Sons, New York, 1987).
21. Agrawal, A. & Schatz, D. G. RAG1 and RAG2 form a stable postcleavage synaptic complex with DNA containing signal ends in V(D)J recombination. *Cell* **89**, 43–53 (1997).
22. Harlow, E. & Lane, D. *Antibodies: A Laboratory Manual* (Cold Spring Harbor Laboratory Press, Plainview, New York, 1989).
23. Voitenleitner, C. & Botchan, M. E1 protein of bovine papillomavirus type 1 interferes with E2 protein-mediated tethering of the viral DNA to mitotic chromosomes. *J. Virol.* **76**, 3440–3451 (2002).
24. Rio, D. C. & Rubin, G. M. Identification and purification of a *Drosophila* protein that binds to the terminal 31-base-pair inverted repeats of the P transposable element. *Proc. Natl Acad. Sci. USA* **85**, 8929–8933 (1988).
25. Royzman, I., Austin, R. J., Bosco, G., Bell, S. P. & Orr-Weaver, T. L. ORC localization in *Drosophila* follicle cells and the effects of mutations in dE2F and dDP. *Genes Dev.* **13**, 827–840 (1999).
26. Orr-Weaver, T. L. *Drosophila* chorion genes: cracking the eggshell's secrets. *Bioessays* **13**, 97–105 (1991).

Supplementary Information accompanies the paper on Nature's website
(<http://www.nature.com/nature>).

Acknowledgements We thank B. Calvi, J. Tower, J. Kadonaga, D. Remus, P. Lewis and K. Fekete for advice, materials and technical contributions to our experiments. This work was supported by a fellowship from the Cancer Research Fund of the Damon Runyon-Walter Winchell Foundation (E.L.B.), and by the National Institutes of Health (J.S.L. and M.R.B.).

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to M.R.B.
(e-mail: mbotchan@uclink4.berkeley.edu).

The RNA processing exosome is linked to elongating RNA polymerase II in *Drosophila*

Erik D. Andrulis*, Janis Werner*, Arpi Nazarian†, Hediye Erdjument-Bromage†, Paul Tempst† & John T. Lis*

* Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA

† Molecular Biology Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

The RNA polymerase II elongation complex contains several factors that facilitate transcription elongation and catalyse the processing of precursor messenger RNAs (pre-mRNAs)^{1–3}. The conserved elongation factor Spt6 is recruited rapidly and robustly to sites of active transcription^{4,5}. Here we show that *Drosophila* Spt6 (dSpt6) co-purifies with the exosome, a complex of 3' to 5' exoribonucleases that is implicated in the processing of structural RNA and in the degradation of improperly processed pre-mRNA^{6–10}. Immunoprecipitation assays of *Drosophila* nuclear extracts show that the exosome also associates with the elongation factor dSpt5 and RNA polymerase II. *In vivo*, exosome subunits colocalize with dSpt6 at transcriptionally active loci on polytene chromosomes during normal development and are strongly recruited to heat-shock loci on gene induction. At higher resolution, chromatin immunoprecipitation analysis shows that the exosome is recruited to transcriptionally active units of heat-shock genes. These data provide a physical basis for the hypothesis that exosome-mediated pre-mRNA surveillance accompanies transcription elongation.

The production of full-length mRNA transcripts requires the coordinated effort of many factors, including the general elongation protein Spt6, which is essential and conserved. In yeast, Spt6 interacts genetically and biochemically with transcription elongation factors that regulate the processivity of RNA polymerase II (Pol II)¹¹. Spt6 may also modulate chromatin structure during transcription, through direct interactions with histone H3 (ref. 12). dSpt6 is recruited to sites of active transcription and colocalizes with the elongating form of Pol II (refs 4, 5). On the basis of these observations, we considered that dSpt6 might associate functionally with the transcriptional elongation machinery.

We took a biochemical approach to identify the factors with which dSpt6 associates. Full-length dSpt6 was cloned, tagged with an epitope and expressed in a stable *Drosophila* cell line (Fig. 1a). Nuclear extracts containing dSpt6-Flag-His₆ (dSpt6FH) were prepared and subjected to Flag affinity chromatography followed by glycerol gradient sedimentation (Fig. 1b). Several polypeptides co-purified with the tagged dSpt6 (Fig. 1c) and were identified unequivocally. We excised bands from the gel and identified the proteins by a combination of peptide mass fingerprinting using matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (MS) and mass spectrometric sequencing using NanoES triple-quadrupole tandem MS (MS/MS).

We identified nine bands with approximate stoichiometry to dSpt6 that represent homologues of the subunits of the yeast and human exosome (Table 1). An additional band corresponded to untagged dSpt6, consistent with the self-association of Spt6 previously deduced from a large-scale analysis of *Saccharomyces cerevisiae* protein complexes¹³. This large-scale analysis failed to detect an interaction between Spt6 and the exosome, which is not unexpected, as our observations suggested that only fractions of the *Drosophila* Spt6 and exosome are associated physically (Fig. 1e, h, i). Nonetheless, this physical association could withstand

stringent purification conditions.

The exosome is a multisubunit complex of 3' → 5' exoribonucleases that has both nuclear and cytoplasmic functions⁷. Exosome function is evolutionarily conserved, because four of the human exosome genes can complement their cognate mutant yeast alleles¹⁴. Originally identified in yeast as a complex involved in the processing of specific ribosomal RNAs¹⁵, the nuclear exosome also has been shown to be critical for degrading unspliced pre-mRNAs⁸. The yeast nuclear exosome also seems to survey pre-mRNA polyadenylation and mRNA release from transcription foci⁶. These studies have led to the proposal that the yeast exosome contributes to a checkpoint

that monitors proper pre-mRNA processing activities.

We found that the *Drosophila* dSpt6-associated exosome is composed of at least nine exosome subunits, comparable to the yeast and human core exosome complexes, which have at least ten subunits¹⁶. Comparative analysis of the *Drosophila* exosome subunits with the yeast and human exosome subunits showed a high degree of sequence conservation (Table 1). Unexpectedly, dRrp40 was not annotated in the National Center for Biotechnology Institute (NCBI) nonredundant (NR) database, but was found by using MS/MS-derived peptide sequences in a BLASTP (basic local alignment search tool protein) search of the whole *Drosophila*

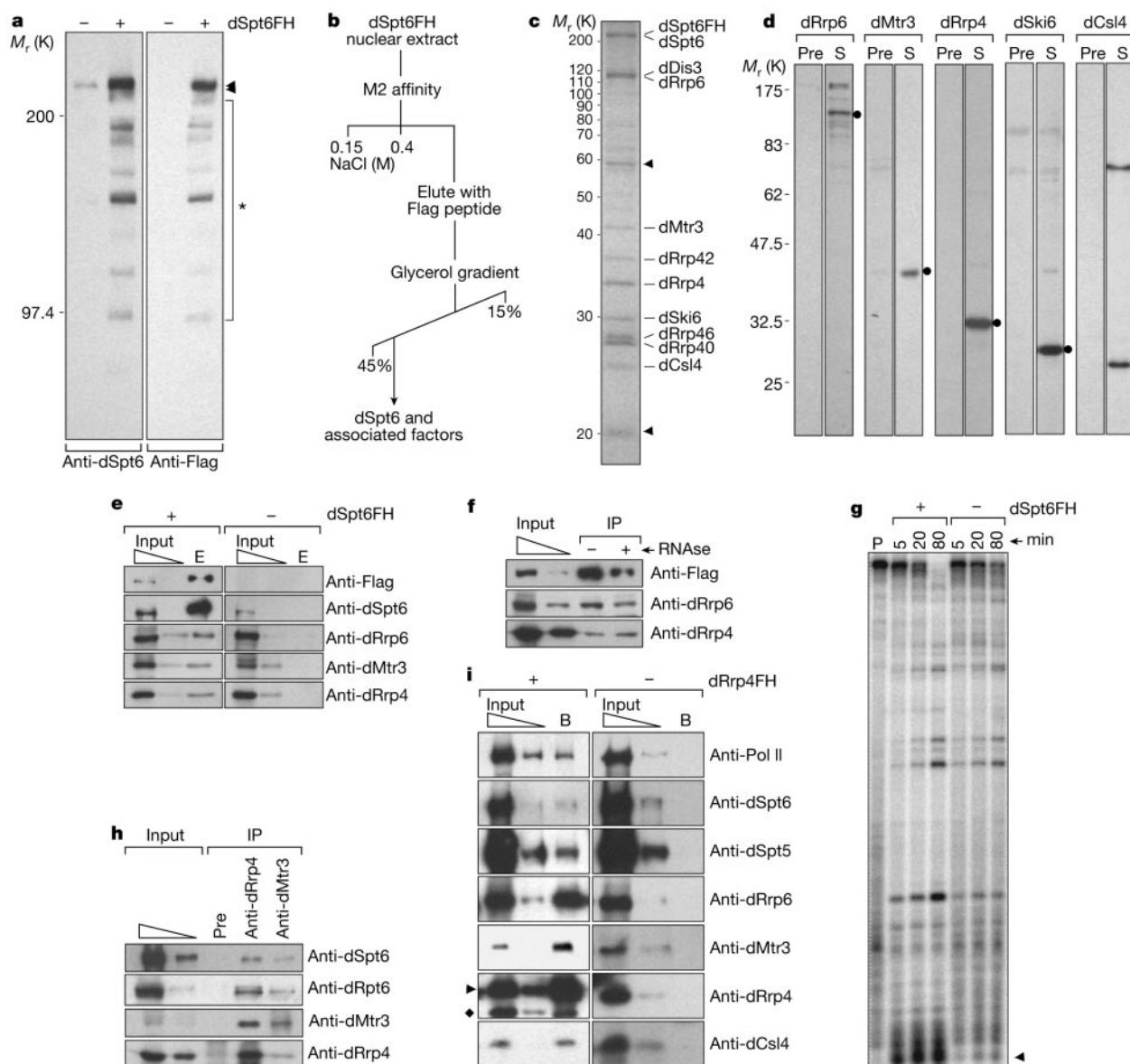


Figure 1 The exosome co-purifies with dSpt6p. **a**, Western blots of whole-cell extracts of stable *Drosophila* Kc cell lines transfected with dSpt6FH (+) or mock transfected (-). Blots were probed with antibodies against Flag or dSpt6. Arrowheads indicate the location of tagged and untagged dSpt6; asterisk indicates N-terminal degradation products of dSpt6FH. **b**, Purification scheme for dSpt6FH and associated factors. **c**, Isolation and identification of the *Drosophila melanogaster* exosome complex. The Coomassie-blue-stained gel shows a glycerol gradient fraction enriched for dSpt6FH-associated factors. Arrowheads and unlabelled bands represent polypeptides that migrate as a peak in a higher fraction of the glycerol gradient (not shown). **d**, Antibody specificity. Western blots of *Drosophila* Kc cell nuclear extracts using antibodies against dRrp6N, dMtr3, dRrp4, dSki6 and dCsl4 (S, serum) and their corresponding

pre-immune serum (Pre). **e**, Exosome purification with an immuno-resin against Flag requires epitope-tagged dSpt6. Western blots of Flag immuno-resin eluate obtained from chromatography with either mock-transfected (-) or dSpt6FH (+) nuclear extracts. Inputs are 5 and 1% of the extract used for immunoprecipitations. **f**, Interaction between the exosome and dSpt6FH is insensitive to RNase. **g**, The dSpt6-exosome complex has exoribonucleolytic activity. Arrowhead indicates the accumulation of single-nucleotide species in the dSpt6FH but not the mock reactions. **h**, Antibodies to exosome subunits specifically immunoprecipitate dSpt6 from *Drosophila* Kc cell nuclear extracts. **i**, Epitope-tagged dRrp4 (dRrp4FH) co-precipitates the exosome, dSpt6, dSpt5 and Pol II. Arrowhead and diamond indicate, respectively, tagged and endogenous dRrp4.

Table 1 *Drosophila* exosome components

Protein	Gene	Human homologue*	Yeast homologue*
dDis3	CG6413	hDis3 53% (69%)	Dis3/Rrp44 52% (67%)
dRrp6	CG7292	PM/Sci-100 34% (52%)	Rrp6 37% (58%)
dMtr3	CG8025	hMtr3 32% (49%)	Mtr3 25% (41%)
dRrp42	CG8395	hRrp42/OIP-2 45% (63%)	Rrp42 25% (50%)
dRrp4	CG3931	hRrp41 57% (77%)	Rrp4 42% (57%)
dSki6	CG15481	hRrp41 53% (70%)	Ski6/Rrp41 32% (56%)
dRrp46	CG4043	hRrp46 38% (54%)	Rrp46 27% (42%)
dRrp40	NA†	hRrp40 43% (67%)	Rrp40 34% (51%)
dCsl4	CG6249	hCsl4 48% (67%)	Cs14 39% (60%)
dRrp45‡	CG9606	PM/Sci-75 36% (60%)	Rrp45 32% (54%)

* Percentages represent sequence identity with sequence similarity given in parentheses.

† NA, not annotated.

‡ Defined by sequence analysis; not present in dSpt6-exosome complex.

genome. The open reading frame (ORF) surrounding the match shared high identity with human Rrp40 and so was named dRrp40.

Notably, the *Drosophila* dSpt6-exosome complex that we purified lacked the Rrp43 and Rrp45 subunits found in the yeast core exosome complex. This might reflect differences in the functional and structural properties of the core and the dSpt6-associated exosome complexes. Although an Rrp45 homologue (dRrp45) is present in the *Drosophila* genome, an Rrp43 homologue is not. This suggests either that *Drosophila* has an unrecognized functional homologue of Rrp43 that did not co-purify with the dSpt6-associated exosome or that Rrp43 does not represent a core exosome component. dRrp45 may be either substoichiometric or absent in the dSpt6-associated exosome complex. The high sequence identity of most subunits indicates, however, that the interactions and functions of the exosome are conserved throughout evolution.

To confirm that the polypeptides co-purifying with dSpt6 represent an exosome complex, we raised antibodies to recombinant proteins corresponding to five of the exosome subunits. Antibodies against dRrp6, dMtr3, dRrp4, dSki6 and dCsl4, but not the pre-immune sera, specifically recognized polypeptides of the expected relative molecular mass (M_r) in nuclear extracts (Fig. 1d). Using these antibodies, we detected dRrp6, dMtr3 and dRrp4 in the Flag immuno-resin eluate from the dSpt6FH extracts but not from control extracts (Fig. 1e). These interactions were insensitive to RNase A (Fig. 1f) and were therefore direct protein-protein interactions.

Because both the purified yeast and human core exosome complexes have exoribonucleolytic activity *in vitro*, we determined whether this dSpt6-exosome complex has the same activity. The Flag immuno-resin eluate containing this complex had 3' → 5'

exoribonucleolytic activity *in vitro* (Fig. 1g), whereas the eluate from the mock-transfected cells had negligible activity. This indicates that the exosome has the capacity to function as an exoribonuclease while in association with the transcription elongation factor dSpt6.

The exosome interaction with dSpt6 could be detected directly in nuclear extracts using exosome-specific antibodies. Antibodies specific for dMtr3 and dRrp4 immunoprecipitated not only other exosome subunits, but also dSpt6 (Fig. 1h). By contrast, pre-immune serum did not immunoprecipitate detectable amounts of dSpt6 or the exosome. To examine whether the exosome associates with additional transcription elongation factors, we analysed immune complexes precipitated from whole-cell extracts of cells transfected with dRrp4-Flag-His₆ (dRrp4FH). We observed that dRrp4FH co-immunoprecipitated roughly 5% of the exosome complex in extracts and 1% of dSpt6 (Fig. 1i), consistent with the dSpt6FH data. We also detected about 1% of the related general transcription elongation factor dSpt5 and the large subunit of Pol II. These results indicate that the exosome associates with an elongation complex in *Drosophila* extracts.

The exosome has been suggested to function at or near sites of transcription⁶. To investigate this, we carried out immunofluorescence studies on *Drosophila* polytene chromosomes with antibodies to exosome subunits. Antibodies against dSki6 stained polytene chromosomes at over 100 loci, including several loci that have high transcriptional activity during development. For example, the developmental puffs at 2B, 74E and 90B showed strong labelling (Fig. 2a). This immunofluorescence staining pattern, like that of many transcription factors, was not static but followed the patterns of gene activation during salivary gland development. The dSki6 and dSpt6 staining patterns showed considerable overlap (Fig. 2a), suggesting that these factors generally colocalize, and perhaps function together, on polytene chromosomes. dCsl4 and dSki6 also colocalize (data not shown). These observations support the hypothesis that dSpt6 and the exosome interact on chromosomes at sites of active transcription, although we cannot rule out that other interactions contribute to this colocalization.

The exosome has been shown to survey heat-shock pre-mRNAs in yeast⁶. To determine whether this surveillance might be occurring co-transcriptionally in *Drosophila*, we tested whether the exosome is recruited to heat-shock loci. Immunofluorescence analysis of dSki6 showed staining at the principal heat-shock loci 67B, 87A, 87C, 93D and 95D (Fig. 2b). In addition, dSki6 was recruited to a transgenic locus containing *hsp70-lacZ*. Antibodies specific for dCsl4 showed an immunofluorescence pattern identical to dSki6 (data not

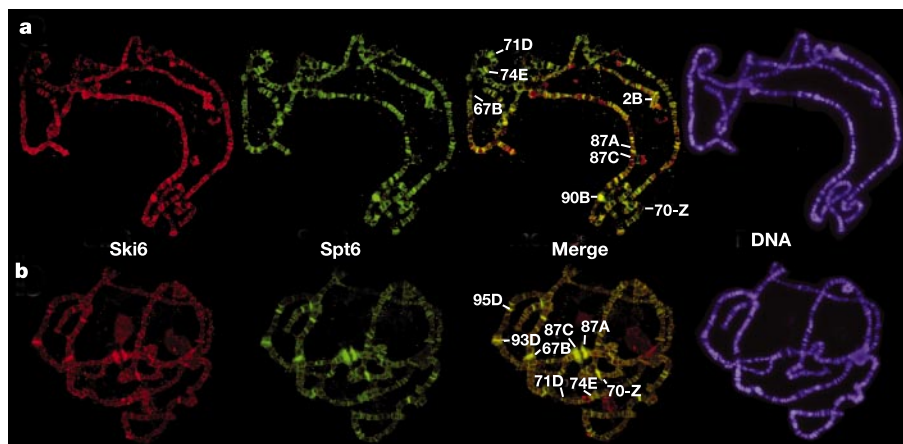


Figure 2 The exosome associates with transcriptionally active loci on polytene chromosomes. Co-immunofluorescence analysis of dSki6 and dSpt6 on salivary gland chromosomes isolated from third instar larva that were either (a) not heat shocked or (b) heat shocked at 37 °C for 20 min. The location of various developmentally active

sites at this puff stage are labelled (2B, 71D, 74E and 90B) as are endogenous heat-shock loci (67B, 87A, 87C, 93D and 95D) and a transgenic heat-shock locus (70-Z, *hsp70-lacZ*)²³. Note the redistribution of both dSki6 and dSpt6 from developmental loci to heat-shock loci on heat shock.

shown). Antibody staining at these endogenous and transgenic loci was coincident with dSpt6 staining. Notably, dSki6 showed a redistribution to heat-shock sites, with a concomitant decrease in staining at non-heat-shock loci. Thus, similar to the general transcription elongation factor dSpt6, the exosome seems to be recruited to and associated with induced genes.

We examined the association of the exosome at higher resolution on heat-shock genes using chromatin immunoprecipitation (ChIP). Formaldehyde-fixed chromatin was isolated from *Drosophila* Kc cells, and protein-DNA complexes were immunoprecipitated using antibodies to heat-shock factor (HSF), dSpt6 and subunits of the exosome. We measured the distribution of these factors along the *hsp70* gene using primer sets that would detect co-immunoprecipitated promoter/5' and co-immunoprecipitated 3' fragments (Fig. 3a). As observed previously⁵, HSF, the transcription factor that binds to the *hsp70* promoter, immunoprecipitated only the 5' end of the gene. By contrast, dSpt6 showed roughly a sixfold recruitment to the body of the *hsp70* gene on gene induction (Fig. 3b). Antibodies specific for the exosome proteins dRrp6, dRrp4 and dCsl4 immunoprecipitated more *hsp70* DNA under heat-shock conditions. The amount of recruitment of exosomal subunits was increased, on average, fivefold on both the 5' and the 3' end of *hsp70* (Fig. 3b). These data are consistent with our immunofluorescence analysis and suggest that the exosome is recruited to the whole body of *hsp70* on induction of this gene.

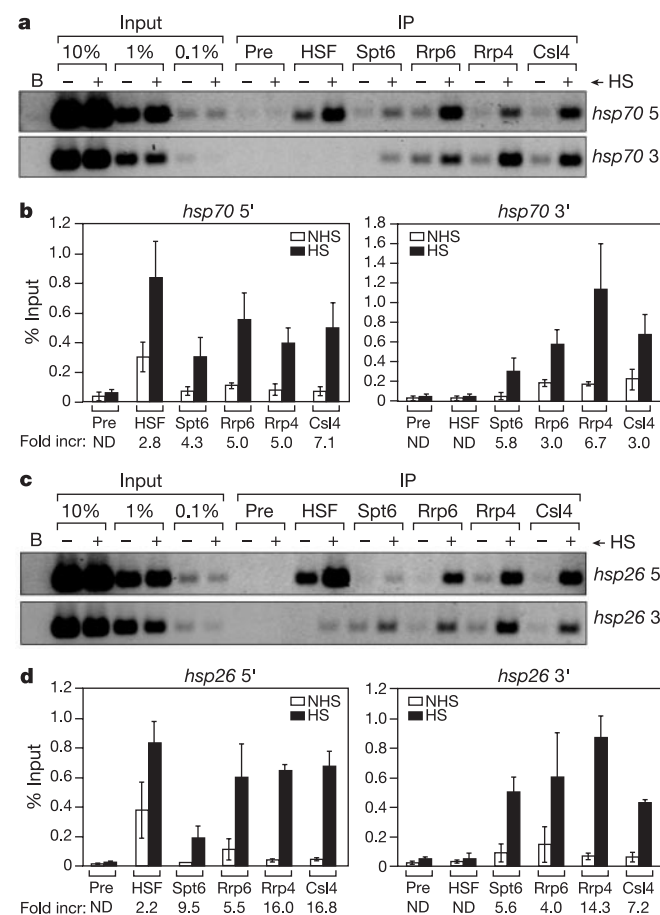


Figure 3 Crosslinking and immunoprecipitation of the exosome at *Drosophila* heat-shock genes. **a, c**, PCR analysis of immunoprecipitated *hsp70* and *hsp26* DNA by electrophoresis on 2% agarose gels stained with ethidium bromide. **b, d**, Quantification of bands stained with ethidium bromide for each antibody presented as a percentage of the total input DNA. B, buffer only control; Fold incr, mean fold increase; HS, heat shock; NHS, no heat shock; ND, not determined; Pre, pre-immune serum.

We also carried out ChIP assays on *hsp26* using the same antibodies (Fig. 3c). We observed similar amounts of recruitment for HSF and roughly an eightfold increase in the recruitment of dSpt6. Quantitative analysis of the recruitment of exosome subunits to *hsp26* showed, on average, an 11-fold increase on heat shock (Fig. 3d). dRrp4 clearly showed the greatest increase at the 3' end of *hsp26* and *hsp70* in material immunoprecipitated from heat-shocked cells. These ChIP data corroborate the immunofluorescence analysis and suggest that the exosome, like dSpt6, tracks with the transcription elongation complex.

Pol II associates with several factors for efficient pre-mRNA processing and elongation¹⁻³. Previous genetic¹¹ and colocalization studies^{4,5} support a direct role for Spt6 in transcription elongation. The interaction and colocalization of dSpt6 and the exosome at transcriptionally active genes indicates that the machinery for transcription elongation and the machinery for pre-mRNA surveillance function together *in vivo*. We propose that the exosome exerts these surveillance functions co-transcriptionally, through an association with Spt6 and elongating Pol II. Together, our observations are consistent with the hypothesis that the exosome surveys and degrades inappropriately made or processed pre-mRNA molecules that may otherwise clog transcription elongation, RNA processing or export pathways. □

Methods

Protein expression in *Drosophila* cells

Drosophila Kc cells were maintained in HyQ-CCM3 media (Hyclone). dSpt6-Flag-His₆ (dSpt6FH) was amplified by polymerase chain reaction (PCR) from LD44036 and LD23412 complementary DNAs (Research Genetics) and fly genomic DNA using the primers Spt6 + 1F and Spt6 + 1831R-Flag-His. Full-length PCR-amplified dSpt6FH was digested with *Sma*I and *Sal*I and cloned into pSK (Stratagene); it was then excised as a *Sma*I-*Sal*I fragment and cloned into pAc5.1/V5-HisA (Invitrogen) that had been digested with *Eco*RV and *Xho*I. We co-transfected Kc cells with Actp-dSpt6FH and pCoHYGRO by using CELLfectin (Invitrogen) and selected stable cell lines with 300 μ g ml⁻¹ Hygromycin B in accordance with the manufacturer's recommendations (Invitrogen). dRrp4FH was amplified by PCR from GM02257 cDNA (Research Genetics) and cloned into pRMHa3 (a gift from P. Mason) for Cu²⁺-inducible expression. We detected the expression of dSpt6FH and dRrp4FH by western blotting using an enhanced chemiluminescence system (Amersham) and either a guinea-pig polyclonal antibody against dSpt6 or an M2 monoclonal antibody against Flag (Sigma).

Nuclear extract preparation and chromatography

Nuclear extracts were prepared as described¹⁷ and incubated with an M2-agarose affinity resin specific for Flag (Sigma) for 2 h at 4 °C before being loaded onto a column. The column was washed with ten bead volumes of buffer V (10 mM Tris, pH 7.5, 0.1 M EDTA, pH 8, 150 mM NaCl, 5 mM MgCl₂, 10% glycerol, 0.1% Nonidet P-40 (NP-40) and protease inhibitors) and with ten bead volumes of buffer VI (buffer V but with 400 mM NaCl) and eluted in buffer V containing 250 μ g ml⁻¹ Flag peptide (Sigma). We collected 50- μ l fractions and analysed them by western blot with antibodies against dSpt6 and by silver staining. Peak fractions were pooled and loaded directly onto a 15–45% glycerol gradient and centrifuged for 16 h at 55,000 r.p.m. (SW60) and 4 °C. We removed 250- μ l fractions from the gradient and analysed them by silver staining and western blotting. The dSpt6-associated complexes migrated as 2–3 peak fractions above the bottom of the gradient; these fractions were precipitated by trichloroacetic acid and loaded onto an SDS-PAGE gel for colloidal Coomassie staining (Novex). The bands were excised for sequencing.

Protein identification and database analyses

Gel-resolved proteins were digested with trypsin and the mixtures were fractionated on a Poros 50 R2 RP micro-tip¹⁸. The resulting peptide pools were then analysed by MALDI reflectron TOF MS using a Reflex III instrument (Bruker Daltonics), and by electrospray ionization (ESI) MS/MS on an API 300 triple quadrupole instrument (Applied Biosystems/MDS SCIEX), modified with an ultrafine ionization source¹⁹. We used selected mass values from the MALDI-TOF experiments to search the NCBI protein NR database using the PeptideSearch²⁰ algorithm. MS/MS spectra from ESI triple quadrupole analyses were inspected for y^+ ion series and the resultant information transferred to the PepFrag²¹ program and used as a search string. Any identification thus obtained was verified by comparing the computer-generated fragment ion series of the predicted tryptic peptide with the experimental MS/MS data. We retrieved *Drosophila* exosome protein sequences from the Berkeley *Drosophila* Genome Database, yeast sequences from the *Saccharomyces* Genome Database, and human sequences from the NCBI. Sequences were compared using BLAST.

Antibody production and use

We cloned full-length exosome genes from cDNAs obtained from Research Genetics: dRrp6 (AT09955 and GM04108), dMtr3 (AT15208), dRrp4 (GM02257), dSki6

(RE67757), and dCsl4 (RE64677). They were amplified by PCR using the following primers: dRrp6 + 1F and dRrp6 + 579R; dMtr3 + 1F and dMtr3 + 326R; Rrp4 + 1F and Rrp4 + 298R; dSki6 + 1F and dSki6 + 246R; dCsl4 + 1F and dCsl4 + 204R. They were digested with the appropriate restriction endonucleases and cloned into pMAL-C2 (New England Biolabs) to create MBP-dRrp6N, MBP-dMtr3, MBP-dRrp4, MBP-dSki6 and MBP-dCsl4. MBP fusions were expressed in *Escherichia coli* and recombinant proteins were purified over amylose resin according to the manufacturer's recommendations (New England Biolabs). Recombinant proteins were injected into either rat or guinea-pig and antibodies were recovered in animal bleeds (Pocono Rabbit Farm and Laboratory Inc.).

We carried out immunoprecipitations in wash buffer (10 mM HEPES (pH 7.5), 10 mM Tris-HCl (pH 7.5), 150 mM KCl, 150 mM NaCl, 3.5 mM MgCl₂, 0.5 mM EDTA, 0.5% NP-40, 10% glycerol, 0.5 mg ml⁻¹ bovine serum albumin, 0.5 mM dithiothreitol (DTT) and 0.25 mM phenyl-methyl-sulphonyl fluoride). RNase A (Sigma) was used at a final concentration of 100 µg ml⁻¹. Immune complexes were incubated with either protein-A- or protein-G-conjugated agarose (Invitrogen) and washed with wash buffer. Beads were boiled with SDS loading dye and then analysed by SDS-PAGE and western blotting.

Exoribonuclease assays

We prepared substrate for *in vitro* exoribonuclease assays as follows. Plasmid pG2XM, containing the 5' end of *hsp70* ORF, was digested with *Eco*RI to linearize the template, and then 1 µg of linearized template was transcribed *in vitro* using the MEGAShortscript kit (Ambion). RNA was 5'-end-labelled by incubating transcription reactions in the presence of [γ -³²P]GTP (Amersham) for 3 min as the only source of GTP. Cold GTP was added thereafter and reactions proceeded for 2 h. We extracted RNA with phenol/chloroform, precipitated it and passed it over a P6 resin (Bio-Rad) to remove unincorporated label. Each exoribonuclease reaction contained ~10 pmol of *hsp70* 5' RNA, ~2 pmol of dSpt6-exosome complex (2 µl of dSpt6FH or mock eluate) and reaction buffer (10 mM Tris, 50 mM KCl, 5 mM MgCl₂ and 10 mM DTT). Reactions were allowed to proceed for the desired time and stopped by the addition of an equal volume of formamide dye. RNA species were separated on a 12% denaturing polyacrylamide gel and analysed by phosphorimaging.

ChIP and immunofluorescence analyses

Crosslinked material was prepared from formaldehyde-fixed Kc cells. We carried out immunoprecipitations with antibodies to *Drosophila* proteins and analysed them as described⁵. Polytenes immunofluorescence was done as described²².

Received 29 July; accepted 16 September 2002; doi:10.1038/nature01181.

- Hirose, Y. & Manley, J. L. RNA polymerase II and the integration of nuclear events. *Genes Dev.* **14**, 1415–1429 (2000).
- Bentley, D. Coupling RNA polymerase II transcription with pre-mRNA processing. *Curr. Opin. Cell Biol.* **11**, 347–351 (1999).
- Proudfoot, N. J., Furger, A. & Dye, M. J. Integrating mRNA processing with transcription. *Cell* **108**, 501–512 (2002).
- Kaplan, C. D., Morris, J. R., Wu, C. & Winston, F. Spt5 and spt6 are associated with active transcription and have characteristics of general elongation factors in *D. melanogaster*. *Genes Dev.* **14**, 2623–2634 (2000).
- Andrulis, E. D., Guzman, E., Doring, P., Werner, J. & Lis, J. T. High-resolution localization of *Drosophila* Spt5 and Spt6 at heat shock genes *in vivo*: roles in promoter proximal pausing and transcription elongation. *Genes Dev.* **14**, 2635–2649 (2000).
- Hillgren, P., McCarthy, T., Rosbash, M., Parker, R. & Jensen, T. H. Quality control of mRNA 3'-end processing is linked to the nuclear exosome. *Nature* **413**, 538–542 (2001).
- Butler, J. S. The yin and yang of the exosome. *Trends Cell Biol.* **12**, 90–96 (2002).
- Bousquet-Antonelli, C., Presutti, C. & Tollervey, D. Identification of a regulated pathway for nuclear pre-mRNA turnover. *Cell* **102**, 765–775 (2000).
- van Hoof, A., Frischmeyer, P. A., Dietz, H. C. & Parker, R. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science* **295**, 2262–2264 (2002).
- Torchet, C. *et al.* Processing of 3'-extended read-through transcripts by the exosome can generate functional mRNAs. *Mol. Cell* **9**, 1285–1296 (2002).
- Hartzog, G. A., Wada, T., Handa, H. & Winston, F. Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in *Saccharomyces cerevisiae*. *Genes Dev.* **12**, 357–369 (1998).
- Bortvin, A. & Winston, F. Evidence that Spt6 controls chromatin structure by a direct interaction with histones. *Science* **272**, 1473–1476 (1996).
- Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415**, 141–147 (2002).
- Allmang, C. *et al.* The yeast exosome and human PM-Scl are related complexes of 3' → 5' exonucleases. *Genes Dev.* **13**, 2148–2158 (1999).
- Mitchell, P., Petfalski, E., Shevchenko, A., Mann, M. & Tollervey, D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3' → 5' exoribonucleases. *Cell* **91**, 457–466 (1997).
- Mitchell, P. & Tollervey, D. Musing on the structural organization of the exosome complex. *Nature Struct. Biol.* **7**, 843–846 (2000).
- Lis, J. T., Mason, P., Peng, J. & Price, D. H. P-TFb kinase recruitment and function at heat shock loci. *Genes Dev.* **14**, 792–803 (2000).
- Erdjument-Bromage, H. *et al.* Examination of micro-tip reversed-phase liquid chromatographic extraction of peptide pools for mass spectrometric analysis. *J. Chromatogr. A* **826**, 167–181 (1998).
- Geromanos, S., Freckleton, G. & Tempst, P. Tuning of an electrospray ionization source for maximum peptide-ion transmission into a mass spectrometer. *Anal. Chem.* **72**, 777–790 (2000).
- Mann, M., Hojrup, P. & Roepstorff, P. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. *Biol. Mass Spectrom.* **22**, 338–345 (1993).
- Fenyo, D., Qin, J. & Chait, B. T. Protein identification using mass spectrometric information. *Electrophoresis* **19**, 998–1005 (1998).

- Park, J. M., Werner, J., Kim, J. M., Lis, J. T. & Kim, Y. J. Mediator, not holoenzyme, is directly recruited to the heat shock promoter by HSF upon heat shock. *Mol. Cell* **8**, 9–19 (2001).
- Shopland, L. S. & Lis, J. T. HSF recruitment and loss at most *Drosophila* heat shock loci is coordinated and depends on proximal promoter sequences. *Chromosoma* **105**, 158–171 (1996).

Acknowledgements We thank members of the Lis laboratory for comments on the manuscript. This work was supported by an NIH grant to J.T.L., a National Research Service Award to E.D.A., and a National Cancer Institute (NCI) Cancer Center Support Grant to P.T.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to J.T.L. (e-mail: jtl10@cornell.edu) or E.D.A. (e-mail: eda4@cornell.edu).

A ribozyme composed of only two different nucleotides

John S. Reader & Gerald F. Joyce

Departments of Chemistry and Molecular Biology and The Skaggs Institute for Chemical Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

RNA molecules are thought to have been prominent in the early history of life on Earth because of their ability both to encode genetic information and to exhibit catalytic function¹. The modern genetic alphabet relies on two sets of complementary base pairs to store genetic information. However, owing to the chemical instability of cytosine, which readily deaminates to uracil², a primitive genetic system composed of the bases A, U, G and C may have been difficult to establish. It has been suggested that the first genetic material instead contained only a single base-pairing unit^{3–7}. Here we show that binary informational macromolecules, containing only two different nucleotide sub-units, can act as catalysts. *In vitro* evolution was used to obtain ligase ribozymes composed of only 2,6-diaminopurine and uracil nucleotides, which catalyse the template-directed joining of two RNA molecules, one bearing a 5'-triphosphate and the other a 3'-hydroxyl. The active conformation of the fastest isolated ribozyme had a catalytic rate that was about 36,000-fold faster than the uncatalysed rate of reaction. This ribozyme is specific for the formation of biologically relevant 3',5'-phosphodiester linkages.

A good starting point for the evolution of a catalyst that contains only two different subunits was the R3 ligase ribozyme, which contains only adenine, guanine and uracil nucleotides^{8,9}. This ribozyme catalyses attack of the 3'-hydroxyl of an RNA substrate on the 5'-triphosphate of the ribozyme, forming a 3',5'-phosphodiester and releasing inorganic pyrophosphate. The chemistry of this reaction is identical to that catalysed by modern RNA polymerase proteins. A templating region within the ribozyme is responsible for binding the RNA substrate, and the sequences of both the template and substrate can be designed such that they contain only adenine and uracil residues. In that format, both the ribozyme and substrate are completely devoid of cytosine and undergo RNA ligation with a catalytic rate, k_{cat} of 0.013 min⁻¹ and Michaelis constant, K_m , of 6.2 µM (ref. 9).

The R3 ribozyme was found to be highly tolerant of base substitutions involving replacement of every adenine by 2,6-diaminopurine (D). Three bonds are formed between 2,6-D and uracil in the context of a Watson-Crick base pair¹⁰, in comparison with adenine, which forms only two. Despite this difference, the D-substituted R3 ligase (Fig. 1a) retained a k_{cat} of 0.001 min⁻¹ and K_m of 12 µM, and reacted to a maximum extent of about 40%.

The second stage of the development of a binary informational catalyst involved *in vitro* evolution to compensate for removal of the final G residues and improve catalytic activity. The sequence of the ribozyme that contained only three G residues was modified by replacing the remaining G residues with either D or U, then introducing random mutations (either $D \rightarrow U$ or $U \rightarrow D$) at a frequency of 12% per nucleotide position. A population of 8×10^{13}

The population of ribozyme variants was given an opportunity to ligate the promoter-containing substrate. The reacted molecules were purified by electrophoresis in a denaturing polyacrylamide gel, then reverse transcribed and PCR amplified in the presence of all four standard deoxynucleoside 5'-triphosphates. In principle, only those molecules that contained D and U residues at positions 1–66 and had catalysed the ligation reaction would be eligible for subsequent transcription to generate progeny molecules. After four rounds of selective amplification, a ligated product was detected in the polyacrylamide gel following the RNA-catalysed reaction. A fifth round was carried out and individual molecules were cloned from the population and sequenced. Only two of the 22 sequenced clones contained a single contaminating G or C residue at positions 1–66. There was considerable sequence heterogeneity among the clones. Only one sequence was found to occur repeatedly, appearing in five of the clones that were examined. Individual

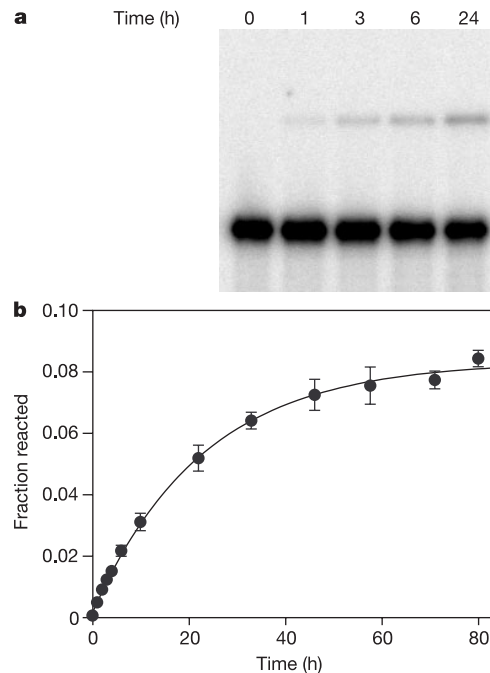


Figure 2 Time course of the cyclization reaction involving the final selected ribozyme, which contained only D and U residues. The reaction mixture contained 20 nM RNA, 100 mM MgCl_2 , 0.01% SDS and 30 mM 2-(*N*-cyclohexylamino)ethanesulphonic acid (CHES; pH 9.0), which was incubated at 23 °C. Products were separated in a 6% denaturing polyacrylamide gel. **a**, Phosphorimager scan of selected time points. **b**, Time course carried out to determine the maximum extent of reaction. Each data point is the average of three separate experiments, with the error bars corresponding to one standard deviation. The data were fitted to a single exponential to obtain a K_{obs} of $0.048 \pm 0.005 \text{ h}^{-1}$ and a maximum extent of reaction of 8.3%.

cloned ribozymes were then prepared in a form that lacked G and C residues within the substrate-binding region. This was accomplished by synthesizing DNA templates based on the sequence of each clone, but with modification of the substrate-binding region so that it contained only D and U residues. The ribozymes containing only D and U residues were then challenged to react with a complementary substrate that contained only A and U residues. All eight of the clones that were tested in this fashion had some detectable activity.

The sequences of the two most active clones were used as the basis to construct two separate pools of RNAs that were mutagenized at a frequency of 12% per nucleotide position. Five additional rounds of selective amplification were performed, resulting in DU-containing ribozymes that were improved with regard to their catalytic activity. Individual clones were again isolated from the final population; the most active of these clones is shown in Fig. 1b. This ribozyme contained many of the mutations that were present in the parent clone, but also contained new mutations, including several reversions. There was also a deletion at the 5' end of the ribozyme that resulted in the 5'-terminal residue being a uridine, located adjacent to an unpaired adenosine at the 3' end of the substrate. Many other mutations were present throughout the final selected ribozyme, presumably resulting in substantial remodelling of its overall structure and the detailed structure at the ligation junction.

The secondary structure of the final selected ribozyme that is depicted in Fig. 1b is drawn by analogy to that of the D-substituted starting molecule, shown in Fig. 1a. Although the secondary structure of the starting molecule is well established⁹, that of the final ribozyme must be regarded as conjectural, especially for the P2 and P3 stems (see Fig. 1a). The final ribozyme can adopt several different structures, as suggested by the presence of multiple bands when it was analysed by non-denaturing polyacrylamide gel electrophoresis (data not shown). As discussed below, most of the molecules were not in an active conformation, which thwarted attempts to carry out meaningful analysis of the ribozyme's secondary structure.

The rate of reaction of the final selected ribozyme was examined in a format in which the ribozyme and substrate contained only D and U residues. Two different constructs were prepared, one in which the substrate was presented separately and another in which it was tethered to the 3' end of the ribozyme by a stable hairpin structure. The latter format allowed for a cyclization reaction

involving the 3'-hydroxyl and 5'-triphosphate of the same RNA, enabling a more straightforward assessment of the fraction of molecules that were in a productive conformation. The observed rate of the cyclization reaction fitted well to a single exponential model, with a k_{obs} of 0.00080 min^{-1} and a maximum extent of about 8% of the total RNA molecules (Fig. 2). A probable explanation for the substantial proportion of unreacted RNA molecules is their propensity to misfold based on their extraordinary degree of internal self-complementarity.

The reaction with a separate RNA substrate was used to show that the ribozyme exhibits Michaelis-Menten saturation kinetics. The ribozyme was engineered to bind a separate 17-nucleotide RNA substrate having the sequence 5'-UUDUUUDDUUDUUDUD-3' (Fig. 1b). Experiments were performed in the presence of excess ribozyme, employing a $[5'-^{32}\text{P}]$ -labelled substrate. On the basis of the initial rates of reaction in the presence of various concentrations of ribozyme, and adjusting for a maximum extent of reaction of 6%, the apparent k_{cat} was 0.0011 min^{-1} and K_m was 1.6 nM (Fig. 3). The uncatalysed rate of ligation was measured under the same reaction conditions employing the same template and substrate sequences. That rate was $3 \times 10^{-8} \text{ min}^{-1}$, which agrees well with previous measurements of uncatalysed template-directed RNA ligation¹², and corresponds to a catalytic rate enhancement of about 36,000-fold. When the RNA-catalysed reaction was performed with a substrate that was not complementary to the template, there was no detectable reaction.

The ligation reaction catalysed by the DU-containing ribozyme may have resulted in the formation of either a 2',5'- or 3',5'-phosphodiester linkage, owing to the lack of selective pressure to maintain the 3',5'-regiospecificity of the starting R3 ligase. The regiospecificity of the reaction was analysed by employing the '10-23' DNA enzyme, which cleaves 3',5', but not 2',5' linkages of RNA¹³. The ligated product was cleaved by the DNA enzyme to generate two fragments of the expected size, demonstrating that the ribozyme, itself composed of 3',5'-linked ribonucleotides, catalyses the formation of a 3',5'-phosphodiester linkage.

It has been suggested that the original genetic system contained only two different nucleotides³⁻⁷, and subsequently evolved to its present, more complex form. A binary genetic system may have been advantageous during the early history of life on earth when the availability of all four nucleotides might have been difficult to maintain. Cytosine nucleotides are especially problematic because they undergo rapid deamination to uridylate, with a half-life of 19 days at pH 7 and 100°C (ref. 2). Thus the G•C pairing may not have been sustainable until the invention of a mechanism for restoring cytosine to uracil. By comparison, the half-life of adenine or diaminopurine at pH 7 and 100°C is about one or two years, respectively². The prebiotic synthesis of adenine or diaminopurine proceeds with comparable efficiency, both compounds being obtained in good yield starting from aqueous ammonium cyanide^{14,15}.

Nucleic acid enzymes composed of only D and U residues pay a heavy price for their simplified composition in terms of both catalytic rate and the fraction of molecules that are in an active conformation. Nonetheless, darwinian evolution can produce catalytically active structures even from such a severely restricted chemical repertoire. The conformational plasticity of DU-containing RNA may even offer an advantage with regard to the ability to explore multiple structures for a given sequence¹⁶.

Ribozymes composed of other pairs of nucleotides, such as A and U, G and C, or even A and I (inosine), may be possible. It seems less likely that polymers composed of only two amino acids could exhibit appreciable catalytic activity. Thus far a minimum of 14 amino acids has been used to construct a catalytic polypeptide¹⁷, and as few as seven have been shown to be necessary to define a folded tertiary structure¹⁸. The absolute minimum number of distinct subunits that could be used to construct a functional

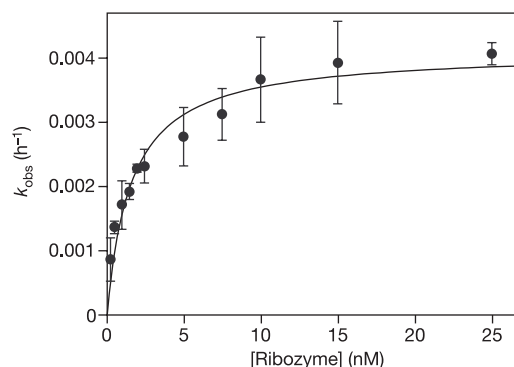


Figure 3 Saturation plot for the bimolecular reaction involving the final selected ribozyme and an RNA substrate, both of which contained only D and U residues. Initial rates were determined for the ligation reaction involving 0.25–25 nM ribozyme and a trace amount of $[5'-^{32}\text{P}]$ -labelled substrate. The reaction mixture contained 100 mM MgCl_2 , 0.01% SDS and 30 mM CHES (pH 9.0), which was incubated at 23°C . Each value for k_{obs} represents the average of three independent experiments, with the error bars corresponding to one standard deviation. The data were fit to a Michaelis-Menten saturation plot to obtain a k_{cat} of $0.0041 \pm 0.0006 \text{ h}^{-1}$ and K_m of $1.6 \pm 0.9 \text{ nM}$. The k_{cat} was adjusted to account for a 6% maximum extent of reaction, giving a value of $0.068 \pm 0.010 \text{ h}^{-1}$.

☐

Synthesis of oligonucleotides

Construction of starting pool and *in vitro* evolution

RNA-catalysed RNA ligation was carried out in the presence of 0.5 μM pool RNA, 5 μM substrate having the sequence 5'-GCCCTCCGAACGCTCTTAATACGACTCACUAGA-3' (T7 RNA polymerase promoter sequence underlined: RNA portion in bold), 25 mM MgCl_2 , 50 mM KCl, 30 mM N-(2-hydroxyethyl)-piperazine-N'-3-propanesulphonic acid (EPPS; pH 8.5), 4 mM dithiothreitol (DTT), and 2 mM spermidine, which were incubated at 23°C for 17 h. It should be noted that the promoter sequence differs from the standard T7 promoter at the next-to-last position, which was found to be beneficial for initiating transcription with a nucleotide other than guanylate¹⁹. The ligated products were separated in a 6% denaturing polyacrylamide gel, eluted from the gel, and precipitated with ethanol in the presence of 20 pmol of a DNA primer having the sequence 5'-GACCGTACCAATCCAGC-3'. The primer was used to initiate reverse transcription in the presence of all four dNTPs. The reverse transcripts were precipitated with ethanol, then PCR-amplified using primers 5'-GACCGTACCAATCCAGC-3' and 5'-GCCCTCCGAACGCTCC-3'. The PCR products were purified using a Qiagen PCR purification kit, then transcribed in the presence of only DTP and UTP to initiate the next round of *in vitro* evolution. Subsequent rounds were performed similarly, except that they were carried out on a smaller scale and employed progressively shorter incubation times during the RNA-catalysed reaction.

The kinetic properties of the ribozyme containing D, G and U residues were measured in the presence of 25 mM MgCl₂, 50 mM NaCl and 25 mM EPPS (pH 8.5) at 23 °C, employing trace amounts of [α -³²P]UTP-labelled ribozyme and excess RNA substrate having the sequence 5'-UUAUAAAAUAUA-3'. A Michaelis-Menten saturation plot was generated on the basis of the initial rates of reaction and correcting for the maximum extent of reaction as determined by long time points. The cyclization reaction involving the final selected ribozyme was carried out employing 20 nM of an RNA that contained both the ribozyme and substrate domains, which was heated to 94 °C for 1 min, then rapidly cooled on ice before initiating the reaction by addition of 100 mM MgCl₂, 0.01% SDS, and 30 mM 2-(*N*-cyclohexylamino)ethanesulphonic acid (CHES; pH 9.0). The products were separated in a 6% denaturing polyacrylamide gel, with a running buffer that contained 40 mM Tris-borate and 0.9 mM Na₂EDTA, then quantified using a phosphorimager. The time course of the reaction was fit to a single exponential, with a maximum extent of 8.3%. The intermolecular reaction involving the final selected ribozyme was carried out under the same conditions as above, except that it employed 0.25–25 nM ribozyme and a trace amount of [5'-³²P]-labelled substrate. The maximum extent of reaction was determined in the presence of 1 μ M ribozyme. A Michaelis-Menten saturation plot was constructed based on the initial rates of reaction and used to obtain values for k_{cat} and K_m .

Analysis of regiospecificity

Received 30 August; accepted 17 September 2002; doi:10.1038/nature01185.

1. Gesteland, R. F., Cech, T. R. & Atkins, J. F. (eds) *The RNA World* 2nd edn (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 1999).
2. Levy, M. & Miller, S. L. The stability of the RNA bases: implications for the origin of life. *Proc. Natl. Acad. Sci. USA* **95**, 7933–7938 (1998).
3. Rich, A. *Horizons in Biochemistry* (eds Kasha, M. & Pullman, B.) 103–126 (Academic, New York, 1962).
4. Crick, F. H. C. The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379 (1968).
5. Orgel, L. E. Evolution of the genetic apparatus. *J. Mol. Biol.* **38**, 381–393 (1968).
6. Wächtershäuser, G. An all-purine precursor of nucleic acids. *Proc. Natl. Acad. Sci. USA* **85**, 1134–1135 (1988).
7. Zubay, G. An all-purine precursor of nucleic acids. *Chemtracts*, **2**, 439–442 (1991).
8. Rogers, J. & Joyce, G. F. A ribozyme that lacks cytidine. *Nature* **402**, 323–325 (1999).
9. Rogers, J. & Joyce, G. F. The effect of cytidine on the structure and function of an RNA ligase ribozyme. *RNA* **7**, 395–404 (2001).
10. Kirnos, M. D., Khudyakov, I. Y., Alexandrushkina, N. I. & Vanyushin, B. F. 2-Aminoadenine is an adenine substituting for a base in 5-2L cyanophage DNA. *Nature* **270**, 369–370 (1977).
11. Macdonald, L. E., Zhou, Y. & McAllister, W. T. Termination and slippage by bacteriophage T7 RNA polymerase. *J. Mol. Biol.* **232**, 1030–1047 (1993).
12. Rohatgi, R., Bartel, D. P. & Szostak, J. W. Kinetic and mechanistic analysis of nonenzymatic, template-directed oligoribonucleotide ligation. *J. Am. Chem. Soc.* **118**, 3332–3339 (1996).
13. Santoro, S. W. & Joyce, G. F. A general purpose RNA-cleaving DNA enzyme. *Proc. Natl. Acad. Sci. USA* **94**, 4262–4266 (1997).
14. Oró, J. Mechanism of synthesis of adenine from hydrogen cyanide under plausible primitive Earth conditions. *Nature* **191**, 1193–1194 (1961).
15. Sanchez, R. A., Ferris, J. P. & Orgel, L. E. Studies in prebiotic synthesis IV. Conversion of 4-aminoimidazole-5-carbonitrile derivatives to purines. *J. Mol. Biol.* **38**, 121–128 (1968).
16. Joyce, G. F. Evolutionary chemistry: getting there from here. *Science* **276**, 1658–1659 (1997).
17. Taylor, S. V., Walter, K. U., Kast, P. & Hilvert, D. Searching sequence space for protein catalysts. *Proc. Natl. Acad. Sci. USA* **98**, 10596–10601 (2001).
18. Plaxco, K. W., Riddle, D. S., Grantcharova, V. & Baker, D. Simplified proteins: minimalist solutions to the 'protein folding problem'. *Curr. Opin. Struct. Biol.* **8**, 80–85 (1998).
19. McGinness, K. E., Wright, M. C. & Joyce, G. F. Continuous *in vitro* evolution of a ribozyme that catalyzes three successive nucleotidyl addition reactions. *Chem. Biol.* **9**, 585–596 (2002).

Acknowledgements We thank J. Rogers for many discussions during the initial stages of this project. We also thank the members of the Joyce laboratory for their advice and E. Tzima for assistance in preparation of the manuscript. This work was supported by a grant from the National Aeronautics and Space Administration and the Skaggs Institute for Chemical Biology. J.S.R. was supported by a postdoctoral fellowship from the NASA Specialized Center for Research and Training (NSCORT) in Exobiology.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to G.F.J. (e-mail: gjoyce@scripps.edu).

nature insight

INFLAMMATION



Cover illustration
HMGB1-expressing cells in invading pannus tissue in collagen-induced arthritis (courtesy of Ulf Andersson).

Inflammation is the response of an organism's immune system to the damage caused to its cells and vascularized tissues by microbial pathogens such as viruses and bacteria, as well as by injurious chemicals or physical insults.

Although painful, inflammation is usually a healing response. But in some instances inflammation proceeds to a chronic state, associated with debilitating disease such as arthritis, multiple sclerosis, or even cancer. At times, acute local inflammation leads to a body-wide response, which can spiral out-of control leading to major organ failure and death.

In this month's *Nature Insight* we bring together a collection of articles exploring how the inflammatory response is set in motion and ultimately controlled. Other articles take a closer look at the adverse role played by inflammation in the aetiology of some of the most prevalent diseases in modern society, and discuss ways in which both acute and chronic inflammatory processes may be amenable to novel methods of therapeutic intervention.

We are pleased to acknowledge the financial support of AstraZeneca in producing this Insight. As always, *Nature* carries sole responsibility for all editorial content and peer-review. We hope that both general readers and experts in the field will find these articles useful and informative.

Ursula Weiss Senior Editor

review articles:

- 846** **Points of control in inflammation**
C. Nathan

- 853** **The inflammatory reflex**
K. J. Tracey

- 860** **Inflammation and cancer**
L. M. Coussens & Z. Werb

- 868** **Inflammation in atherosclerosis**
P. Libby

progress:

- 875** **Mast cells in autoimmune disease**
C. Benoist & D. Mathis

review articles:

- 879** **Inflammation and therapeutic vaccination in CNS diseases**
H. L. Weiner & D. J. Selkoe

- 885** **The immunopathogenesis of sepsis**
J. Cohen

Editor, *Nature*: Philip Campbell
Insights Editors: Lesley Anson
Ursula Weiss
Editorial Assistant: Simon Gibson

Art Director: Majo Xeridat
Diagrams: Ann Thomson
Vicky Askew
Suzanne Coleman

Production Editor: Simon Gribbin
Production: Sue Gray
Marketing: Claire Aspinall
Sponsorship: Mark Greene

Points of control in inflammation

Carl Nathan

Department of Microbiology and Immunology and Graduate Programs in Immunology and Molecular Biology, Weill Medical College of Cornell University, Box 62, 1300 York Avenue, New York 10021, USA (e-mail: cnathan@med.cornell.edu)

Inflammation is a complex set of interactions among soluble factors and cells that can arise in any tissue in response to traumatic, infectious, post-ischaemic, toxic or autoimmune injury. The process normally leads to recovery from infection and to healing. However, if targeted destruction and assisted repair are not properly phased, inflammation can lead to persistent tissue damage by leukocytes, lymphocytes or collagen. Inflammation may be considered in terms of its checkpoints, where binary or higher-order signals drive each commitment to escalate, go signals trigger stop signals, and molecules responsible for mediating the inflammatory response also suppress it, depending on timing and context. The non-inflammatory state does not arise passively from an absence of inflammatory stimuli; rather, maintenance of health requires the positive actions of specific gene products to suppress reactions to potentially inflammatory stimuli that do not warrant a full response.

The 'inflammatory process'¹ includes a tissue-based startle reaction to trauma; go/no-go decisions based on integration of molecular clues for tissue penetration by microbes; the beckoning, instruction and dispatch of cells; the killing of microbes and host cells they infect; liquefaction of surrounding tissue to prevent microbial metastasis; and the healing of tissues damaged by trauma or by the host's response. If at any step an order to proceed is issued but progress to the next step is blocked, the inflammatory process may detour into a holding pattern, such as infiltration of a tissue with aggregates of lymphocytes and leukocytes (granulomas) that are sometimes embedded in proliferating synovial fibroblasts (pannus), or distortion of a tissue with collagen bundles (fibrosis). Persistent inflammation can oxidize DNA badly enough to promote neoplastic transformation.

What Celsus defined around AD40 as 'rubor, calor, dolor, tumor' (redness, heat, pain and swelling) is today an intellectually engaging problem in signal transduction and systems biology, as well as a multibillion dollar market for the pharmaceutical industry. When primary pathogenetic events are unknown, control of inflammation is sometimes the next best option. The number of diseases considered 'inflammatory' in origin may decline as infectious causes continue to be discovered for some of them, such as *Helicobacter pylori*-dependent chronic gastritis with ulcer formation. However, in this and several other important infectious diseases, the inflammatory response may cause more damage than the microbe. Although the search continues for possible infectious causes of multiple sclerosis, rheumatoid arthritis and atherosclerosis, inflammation *per se* remains one of the main therapeutic targets in diverse disorders with a staggering collective impact (Table 1).

Inflammation is usually life preserving, as reflected by the increased risk of grave infections in people with genetic deficiencies in principal components of the inflammatory process. For example, inability to mobilize leukocytes to sites of inflammation in type I or II leukocyte adhesion deficiency, if untreated, often leads to death from infection². Inability to produce the complement components properdin and factors D, C5, C6, C7, C8 or C9 predisposes to meningococcal infection³. Thus, the medical focus on inhibiting inflammation is accompanied by an effort of potentially comparable importance to learn how to induce inflammation more effectively,

in at least two important settings. First, causing and prolonging inflammation are among the essential functions of adjuvants, and a better understanding of the role of inflammation in adjuvanticity may enable prophylactic immunization against a wider range of infectious diseases. Second, generation of inflammation is one of the main goals of tumour immunology, both for therapeutic immunization⁴ and for nonspecific immunostimulation, such as by instilling Bacille Calmette-Guérin into the urinary bladder to prevent recurrence of tumours⁵.

The accompanying articles in this issue integrate cross-sections of inflammation biology by peering inside blood vessels, joints, brain, viscera and epithelia. The papers form a backdrop against which to evaluate diverse new anti-inflammatory treatments. These include neutralizers of tumour-necrosis factor (TNF); blockers of leukotriene receptors; inhibitors of cyclooxygenase (COX)-2, leukotriene synthetase and 3-hydroxy-3-methylglutaryl coenzyme A reductase; and agonism at protease-activated receptor 1 by activated protein C (ref. 6). Many more anti-inflammatory compounds are in the pipeline.

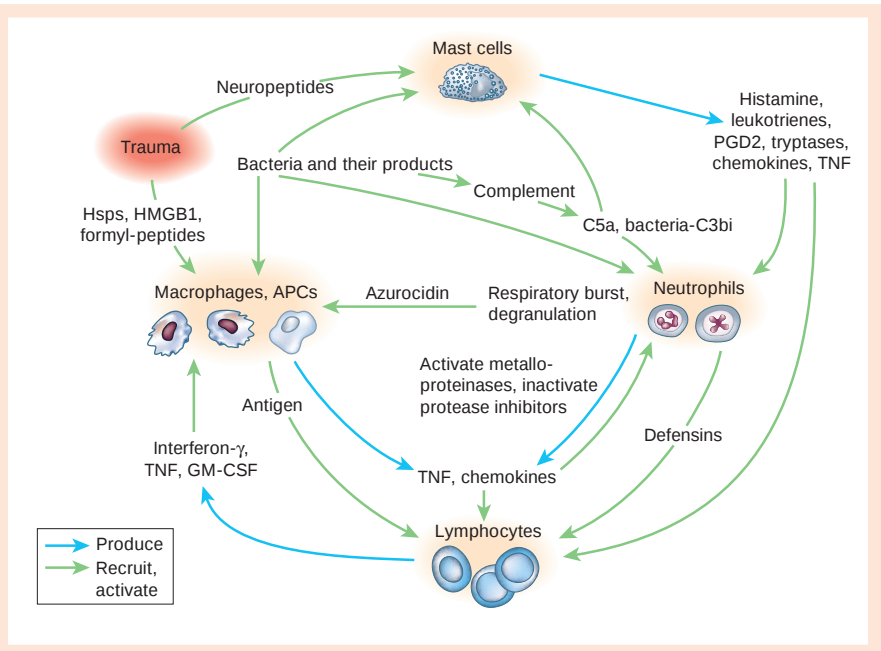
In this article I offer a perspective on inflammation as a system of information flow in response to injury and infection. If tissue is injured, the basic challenge facing the host is to detect whether there is accompanying infection. If infection is the initial event, the challenge is to detect whether tissue is injured. When injury and infection coincide, the goal is to react as quickly as possible to terminate the spread of infection, even at the cost of further tissue damage. The need to detect two states at once before risking self-inflicted damage dictates a dependence on binary or higher-order signals. The need to accelerate at a potentially high cost brings with it the need to decelerate as soon as the goal has been met. A full stop requires repairing the tissue whose damage triggered inflammation or that inflammation damaged.

Such a complex system can be characterized by its checkpoints. I first consider checkpoints evident early and late after an inflammatory response is activated, and then present evidence that another set of checkpoints operates constitutively in the basal state to prevent the inappropriate initiation of inflammation.

Go signals in early checkpoints

Evolution did not anticipate surgery with aseptic technique. Thus, the body reacts to trauma as if the emergency is infec-

Figure 1 Information flow in the early stages following mild trauma with infection. Each cell commits to recruit and activate others based on multiple inputs, generally requiring evidence of both injury and infection, before it joins fully in amplifying the inflammatory process. Not shown are interactions among leukocytes, endothelium, platelets and coagulation factors; the generation of stop signals; and the flow of information over subsequent days, including the transition to wound healing (see text).



tion, until proven otherwise. For simplicity, the present discussion deals with mild trauma and considers only some of the go signals.

The take-home message is apparent with the following experiment. Expose one forearm with the inner surface facing up. Spread the three middle fingers on your other hand and slap them down hard on your forearm. Within about 15 seconds the skin of your forearm will display a red bas-relief of the offending digits. Over the next hour the image will fade. In contrast, if the epidermis had been broken and bacteria had entered, redness and swelling would persist, testifying to an escalating series of events that is synchronized according to bacterial replication time and metastatic potential. The episode would probably culminate in the confinement and killing of the penetrant bacteria and the destruction and repair of a small amount of tissue. Then again, if the inflammatory response were feeble and antibiotics unavailable, the outcome might be death from sepsis.

Figure 1 schematizes the flow of information following mild trauma with infection. Tissue damage unleashes up to three types of go signals. First, in response to pain, neurons release bioactive peptides⁷. Second, broken cells release constitutively expressed intracellular proteins that trigger cytokine production when found in the extracellular space. Examples include heat-shock proteins⁸, the transcription factor HMGB1 (for high mobility group 1)⁹ and mitochondrial peptides bearing the *N*-formyl group characteristic of prokaryotic proteins¹⁰. Third, microbes and their shed or secreted products are sensed through binding of their conserved molecular constituents to soluble receptors such as complement, mannose-binding protein and lipopolysaccharide-binding protein, and to cell-surface receptors such as Toll family members, peptidoglycan recognition proteins and scavenger receptors.

Much attention in inflammation research has focused on the recruitment of leukocytes from the blood¹¹. However, a rapid response requires sentinel cells pre-stationed in the tissues. Mast cells and macrophages fulfil this function. The importance of mast cells as first responders (see review in this issue by Benoist and Mathis, pages 875–878), recently emphasized in experimental rheumatoid arthritis¹², is symbolized by their placement atop Fig. 1. Responding to the signals listed above, perivascular mast cells release histamine, eicosanoids, pre-formed TNF, newly synthesized cytokines, tryptases, other proteases, and chemokines. Histamine, eicosanoids and tryptases cause vasodilatation (responsible for the heat and redness) and extravasation of fluid (the cause of swelling).

Mast-cell tryptases cleave protease-activated receptors whose neo-termini then engage G-protein-coupled receptors on mast cells, sensory nerve endings⁷, endothelium and neutrophils. This further activates mast cells and neurons, makes endothelium sticky for leukocytes and leaky to fluid, and prompts leukocytes to release platelet-activating factor (PAF). PAF reinforces the pro-adhesive conversion of endothelium, which results in leukocyte emigration from the vasculature. For simplicity, interactions among endothelial cells, leukocytes and extravascular signals¹¹ are omitted from Fig. 1. Also omitted here, but discussed in this issue by Cohen (pages 885–891), are the impacts of the coagulation and kinin cascades on interactions of endothelium and leukocytes^{6,13,14} and the reciprocal influence of inflammation on the interactions of endothelium and coagulation factors (see review in this issue by Libby, pages 868–874).

Neutrophils are partially activated (primed) by the TNF and leukotrienes produced by mast cells and by other neutrophils,

Table 1 Examples of inflammatory disorders

Disorders in which an important pathogenetic role is assigned to inflammation	
Alzheimer's disease	Osteoarthritis
Anaphylaxis	Pemphigus
Ankylosing spondylitis	Periodic fever syndromes
Asthma	Psoriasis
Atherosclerosis	Rheumatoid arthritis
Atopic dermatitis	Sarcoidosis
Chronic obstructive pulmonary disease	Systemic lupus erythematosus
Crohn's disease (regional enteritis)	Type I diabetes mellitus
Gout	Ulcerative colitis
Hashimoto's thyroiditis	Vasculitides (Wegener's syndrome, Goodpasture's syndrome, giant cell arteritis, polyarteritis nodosa)
Ischaemia-reperfusion injury (occlusive and embolic stroke and myocardial infarction)	Xenograft rejection
Multiple sclerosis	
Diseases of infectious origin in which inflammation may contribute as much to pathology as does microbial toxicity	
Bacterial dysentery	Influenza virus pneumonia
Chagas disease (<i>Trypanosoma cruzi</i>)	Leprosy (tuberculoid form)
Cystic fibrosis pneumonitis	Neisserial or pneumococcal meningitis
Filariasis	Post-streptococcal glomerulonephritis
<i>Helicobacter pylori</i> gastritis	Sepsis syndrome
Hepatitis C	Tuberculosis
Diseases of diverse origin in which post-inflammatory fibrosis is a principal cause of pathology	
Bleomycin-induced pulmonary fibrosis	Hepatic cirrhosis (post-viral or alcoholic)
Chronic allograft rejection	Radiation-induced pulmonary fibrosis
Idiopathic pulmonary fibrosis	Schistosomiasis

Table 2 Products encoded by genes whose disruption or mutation leads to spontaneous inflammation*

Gene product(s)	Human (H) or mouse (M)	Inflammatory phenotype	Predominant sites	Reference
Factors directly involved in regulation of apoptosis				
Fas (CD95)	H	Urticarial rash, glomerulonephritis, oral ulceration, lymphocyte infiltration	Skin, kidney, mouth, liver	50–52
Fas (CD95) (lpr)	M	Glomerulonephritis, necrotizing vasculitis, erosive synovitis, interstitial pneumonitis, dermatitis	Kidney, mesentery, joints, lung, skin, vessels	53
Factors thought to be involved in clearance of immune complexes and material from apoptotic cells				
C1q (A, B and C genes)	H	Rash, glomerulonephritis, oral ulceration	Skin, kidney, mouth	42
C1q (a gene)	M	Glomerulonephritis	Kidney	54
C2	H	Rash, vasculitis, arthritis, glomerulonephritis, asthma	Skin, joints, lung	42, 55
C4	H	Rash	Skin	43
C4	M	Glomerulonephritis	Kidney	56
C3	H	Glomerulonephritis	Kidney	43
C4-binding protein	H	Ulcerations	Mouth	57
Factor H	H	Glomerulonephritis, rash	Kidney, skin	57
Cry	M	Neutrophil infiltration	Placenta	58
Serum amyloid P component	M	Glomerulonephritis	Kidney	59
DNAse I	M	Glomerulonephritis	Kidney	60
FCγRIIB	M	Glomerulonephritis	Kidney	61
WASP (Wiskott–Aldrich syndrome protein)	H	Eczema, vasculitis, renal disease, arthritis, inflammatory bowel disease	Skin, kidney, joints, bowels	62, 63
WASP	M	Lymphocyte and neutrophil infiltration	Colon	64
Cytokines, cytokine receptors and other cell surface receptors				
TNF-R1 (tumour-necrosis factor receptor 1)	H	Familial Hibernian fever (periodic fever, conjunctivitis, periorbital edema, arthralgia)	Systemic, eyes, joints	65
TGF-β1 (transforming growth factor-β1)	M	Macrophage, lymphocyte and neutrophil infiltration in blood vessels and parenchyma; gastric ulceration	Lung, heart, stomach, liver spleen, lymph nodes, pancreas, colon, salivary glands, striated muscle	66, 67
IL-2Rα (interleukin-2 receptor-α)	M	Lymphocyte and neutrophil infiltration; ulceration	Colon	68
IL-2	M	Granulocyte, lymphocyte and plasma cell infiltration; ulceration	Colon	69
IL-10	M	Lymphocyte and neutrophil infiltration	Duodenum, jejunum, ileum, colon	70
GM-CSF (granulocyte–macrophage colony-stimulating factor)	M	Lymphocyte infiltration around airways and veins	Lung	71
IL-1Ra (IL-1 receptor antagonist)	M	Neutrophil, macrophage and CD4 ⁺ lymphocyte infiltration	Aorta, coronaries, iliac and popliteal arteries	72
IL-1Ra	M	Erosive arthritis	Joints	73
T-cell receptor-α	M	γδ T-cell, B-cell, plasma cell and neutrophil infiltration	Colon	74
T-cell receptor-β	M	γδ T-cell, B-cell, plasma cell and neutrophil infiltration	Colon	74
Major histocompatibility complex class II	M	Lymphocyte and neutrophil infiltration	Colon	74
CTLA4 (cytotoxic T-lymphocyte antigen 4)	M	Lymphocyte, macrophage and granulocyte infiltration	Heart, pancreas, lung, bone marrow, liver, salivary glands, joints, blood vessels	75, 76
PD-1 (immunoreceptor tyrosine-based inhibitory motif (ITIM)-bearing Ig superfamily member; orphan receptor)	M	Arthritis, glomerulonephritis, carditis	Joints, kidneys, heart	77
Intracellular factors in lymphocytes, leukocytes and epithelial cells affecting their activation				
LAT (linker for activation of T cells)	M	CD4 ⁺ T-cell, eosinophil, B-cell and macrophage infiltration	Multiple organs	78
SOCS1 (suppressor of cytokine signalling)	M	Macrophage infiltration	Liver, lungs, pancreas, heart, skin	79
Phosphatidylinositol 3-phosphate kinase p110δ	M	Leukocyte infiltration	Caecum, rectum	80

leading to release of small amounts of elastase. This cleaves the anti-adhesive coat of CD43 (leukosialin) from neutrophils, allowing their integrins to engage extracellular matrix proteins¹⁵. The binary signal of integrin engagement plus stimulation by TNF, chemokines or C5a triggers degranulation and a massive respiratory burst¹⁶, resulting in release of proteinases (such as the serprocidins elastase, cathepsin G and protease 3), other hydrolases, antibiotic proteins (such as bacterial permeability increasing factor, four α-defensins, the three serprocidins and their proteolytically inactive homologue, azurocidin) and oxidants (such as hydrogen peroxide, hypohalites and chloramines). The oxidants activate matrix metalloproteinases (MMPs) and inactivate protease inhibitors¹⁷.

The foregoing actions promote tissue breakdown. Metalloproteinases cleave TNF from tissue macrophages as well as from monocytes that are chemotactically attracted from the bloodstream into the tissue by azurocidin¹⁸. Macrophage- and monocyte-derived TNF and chemokines attract and activate more neutrophils. TNF and

chemokines combine with mast cell-derived prostaglandin E2 (PGE2) and neutrophil-derived defensins to recruit lymphocytes¹⁹, while leukotrienes help attract antigen-presenting dendritic cells²⁰. Lymphocytes, in conjunction with microbial products, activate macrophages to secrete proteases, eicosanoids, cytokines and reactive oxygen and nitrogen intermediates (ROIs and RNIs, respectively).

In summary, the inflammatory system is geared for lag-free acceleration, but requires ongoing verification of emergency to avoid defaulting to the resting state. Each newly recruited cell generally commits to release pro-inflammatory signals only after integrating inputs of both host and microbial origin.

It is a canon of immunology that for cellular activation, B cells generally need antigen-receptor engagement plus signals from T cells; T cells need antigen-receptor engagement plus signals from antigen-presenting cells (APCs); and APCs, including macrophages, need cytokines plus microbial products, or cytokines plus CD40 ligation, or microbial products plus products of necrotic host cells. The

Table 2 Products encoded by genes whose disruption or mutation leads to spontaneous inflammation* (Continued)

Gene product(s)	Human (H) or mouse (M)	Inflammatory phenotype	Predominant sites	Reference
Pten (phosphatase active on PtdIns(3,4,5)P ₃)	M	Inflammatory interstitial infiltration	Lung	81
Lyn	M	Glomerulonephritis, renal capillary vasculitis	Kidney	82, 83
Cbl	M	Activated B- and T-cell infiltration	Salivary glands, pancreas, liver, intestine, lung, kidney, heart, skeletal muscle, urinary bladder	84
Gα ₁₂	M	Lymphocyte, plasma cell and neutrophil infiltration; ulceration	Colon	85
SHP-1 (protein tyrosine phosphatase)	M	Neutrophil abscesses, interstitial pneumonitis	Skin, lung	86, 87
SHIP (inositol-5-phosphatase)	M	Macrophage infiltration	Lungs	88
p21	M	Leukocyte infiltration, glomerulonephritis	Kidney	89
Tristetraprolin	M	Neutrophil, macrophage and lymphocyte infiltration; pannus formation; bone erosion; glomerulonephritis	Skin, conjunctivae, joints, kidney	90
NFAT (nuclear factor of activated T cells)-p and NFAT4 (double deletion)	M	Lymphocyte, macrophage, plasma cell, neutrophil, and mast cell infiltration	Eyelids, lungs	91
IKK (κB kinase)-2 (deficiency restricted to keratinocytes)	M	Macrophage, neutrophil and CD4 ⁺ T-cell infiltration	Skin	92
IKKγ (NEMO: NF-κB essential modulator)	H	Rash; eosinophil and neutrophil infiltration	Skin	93
IKKγ (NEMO: NF-κB essential modulator)	M	Neutrophil and eosinophil infiltration; iNOS expression	Skin	94, 95
IκBα (inhibitor of NF-κB)	M	Neutrophil abscesses and macrophage infiltration	Skin	96
E3 ubiquitin ligase	M	Macrophage and plasma cell infiltration; fibrosis; alveolar proteinosis	Lung	97
RelB	M	T-cell, eosinophil, neutrophil, macrophage and mast cell infiltration; collagen deposition; hyperplasia of mucus-secreting cells	Skin, lungs, liver, salivary glands, skeletal muscles, stomach, epididymis, ovaries, uterus	98–100
NF-κB1 (p105) (retaining p50)	M	Lymphocyte infiltration around vessels, portal tracts and airways	Lungs, liver	101
T-bet	M	Peribronchial and perivascular eosinophil and lymphocyte infiltration; collagen deposition	Lungs	102
Gadd45a (growth arrest and DNA damage-inducible gene)	M	Glomerulonephritis, perivascular mononuclear infiltration	Kidney	103
Mdr (multiple drug resistance)-1a	M	CD4 ⁺ T-cell, B-cell and granulocyte infiltration; mucosal thickening and ulceration	Colon	104
NOD2/CARD15	H	Crohn's disease (granulomas, fibrosis)	Ileum	105, 106
NOD2/CARD15	H	Blau syndrome (granulomatous arthritis, uveitis, rash)	Joints, eyes, skin	107
Pyrin	H	Familial Mediterranean fever (fever, neutrophil infiltration)	Systemic, joints, peritoneum, pleural space	108, 109
Cryopyrin or CIAS (cold-induced autoinflammatory syndrome)-1	H	Cold-induced periodic fever, rash, arthralgia, conjunctivitis; or without cold induction (Muckle-Wells syndrome)	Systemic, skin, joints, conjunctivae	110
Mevalonate kinase	H	Periodic fever, arthralgia, abdominal pain, rash	Systemic, joints, peritoneum (?), skin	111, 112
Factors affecting oxidative stress				
Haem oxygenase 1	M	Lymphocyte, neutrophil and macrophage infiltration; fibrosis; glomerulonephritis	Liver, lung, kidney	113
Surfactant protein D	M	Monocyte peribronchiolar and perivascular infiltration; emphysema	Lung	44

*"Spontaneous" indicates substantially more frequent occurrence than in wild-type hosts, despite lack of known infection or experimental intervention, in subjects living under normal conditions, presumably with normal gastrointestinal flora. People with Wiskott-Aldrich syndrome experience frequent infections but the disorder is included for comparison with the mutant mouse, in which infections have not been described. Only inflammatory aspects of the phenotype are listed, even when these are not the major manifestations of the disorder. Abnormal accumulation of lymphocytes in lymphoid organs and the development of autoantibodies are not considered signs of inflammation. Entries are omitted for disorders in which the mutated gene has not been identified; the mutation is thought to confer a gain of function; the phenotype requires that two known genes both be mutated; or only a single case has been reported. For mice, the phenotype is described in the strain background in which it is most pronounced, provided that the background does not furnish another known mutation with a related phenotype.

discussion above stresses that a requirement for binary or higher-order go signals begins with the activation of mast cells and neutrophils, and that sustained activation of mast cells and neutrophils usually precedes and conditions the activation of APCs, T cells and B cells as the inflammatory response evolves into the immune response. That a combination of tissue injury plus infection sustains inflammation helps clarify what provokes an immune response^{21,22}.

Massive trauma, post-ischaemic or toxic necrosis, and haemorrhage and resuscitation can each trigger an inflammatory response that appears to be independent of infection. This may reflect the ability of some host cell products that are altered (for example, fragmented matrix proteins or oxidized lipoproteins), abnormally released (for example, heat-shock proteins) or released in abnormally large amounts to interact with receptors (for example, Toll-like receptor 4) that otherwise detect microbial signals²³. Alternatively, cryptic microbial signals may be involved, because

such stresses may be associated with the translocation of bacteria or diffusion of their products across the intestinal wall²⁴.

Stop signals in early checkpoints

Superimposed on the feed-forward cycles illustrated above are sets of brakes. Brakes involving lipid autacoids illustrate one mechanism: to progressively raise the threshold for continuing the inflammatory reaction²⁵. Neutrophil-derived arachidonate serves as substrate for neutrophil 5-lipoxygenase to generate the inflammatory leukotriene B₄. However, as neutrophils infiltrate tissues, they also pass arachidonate to tissue cells expressing 15-lipoxygenase, which produces lipoxins. Lipoxins are a class of oxidized eicosanoids that bind cellular receptors and block neutrophil influx²⁵. Neutrophils also pass to other cells a 5-lipoxygenase intermediate, leukotriene A₄; 15-lipoxygenase converts this to a lipoxin as well²⁵. In this manner, cell-cell interactions favour a transition in the profile of arachidonate products from pro-inflammatory leukotrienes to anti-inflammatory lipoxins. At the

same time, COX2 is induced in macrophages by microbial products and cytokines. COX2 converts arachidonate to PGE2, which contributes to fluid leak from blood vessels. However, as PGE2 levels rise, PGE2 feeds back to inhibit COX2 as well as 5-lipoxygenase, while transcriptionally inducing 15-lipoxygenase in neutrophils. These delayed effects shift arachidonate metabolism towards lipoxin formation in neutrophils themselves²⁵. In this way, over several hours, PGE2, at first a go signal, becomes a stop signal. The anti-inflammatory drug aspirin recapitulates this phenomenon by acetylating COX2; the acetylated enzyme switches from making PGE2 to making lipoxins²⁵.

Studies with gene-disrupted mice highlight additional stop signals. Mice lacking the ectonucleotidase CD39 over-react to chemical irritation of the skin²⁶. Mice deprived of purinergic A2a receptors succumb to normally sublethal doses of microbial and chemical toxins²⁷. These observations suggest that CD39 breaks down extracellular ATP and ADP secreted by activated cells or leaking from broken cells, generating adenosine. Adenosine then acts to suppress inflammatory responses by neighbouring cells.

In another set of examples, mice lacking the cell-surface immunoglobulin superfamily molecule CD200 suffer more macrophage influx and worse experimental autoimmune encephalomyelitis and collagen-induced arthritis than do wild-type mice²⁸. Similarly, mast cells lacking the integrin-binding receptor gp49B1 degranulate excessively in response to immunoglobulin E-antigen complexes²⁹. These studies hint at a wide array of protein-protein interactions among cells, and between cells and their matrix, that temper inflammation in its early phase.

A fourth type of stop signal is issued by the autonomic nervous system. As reviewed by Tracey in this issue (pages 853–859), cholinergic discharge blocks the release of TNF from macrophages in the viscera.

Signals for switching from killing to healing

A crucial commitment made late in inflammation is to convert the response from the antibacterial, tissue-damaging mode to a mode that promotes tissue repair and epithelial closure. The timing is critical — to close a wound before it is disinfected invites disaster. Some of the signals involved are revealed by the failure of mice to resolve late-phase inflammation when they are deficient in the CD44 hyaluronan receptor³⁰, secretory leukocyte protease inhibitor (SLPI)³¹ or TNF^{32,33}. The scenario below integrates findings from these reports; space limitation precludes citing additional examples.

Continuing from the point reached in the description of Fig. 1, as long as microbial and host pro-inflammatory stimuli predominate, macrophage-derived chemokines continue to attract neutrophils. ROI and hyaluronidase from macrophages and neutrophils break down hyaluronan acid in the extracellular matrix to low molecular weight fragments. Like the heat-shock proteins, HMGB1 protein and *N*-formyl peptides described earlier, hyaluronan fragments act as signals of injury, working via CD44 on macrophages to induce the further release of chemokines and perhaps MMPs. Neutrophils with engaged integrins are activated by macrophage-derived TNF to release abundant elastase. Elastase and ROI activate MMPs. MMPs activate macrophage-derived latent transforming growth factor- β (TGF- β), the most potent known chemoattractant for neutrophils. MMPs also degrade collagen, proteoglycans and fibronectin. Elastase degrades latent TGF- β -binding protein, contributing to the activation of TGF- β .

The transformation from tissue damage to tissue repair begins as complement, neutrophils and macrophages kill microbes, and macrophages secrete more SLPI, a serine protease inhibitor expressed late after exposure to microbial products or cytokines. SLPI has anti-inflammatory³⁴ and wound-healing effects³¹ that include suppressing the release of elastase and ROI by TNF-stimulated neutrophils³⁵, inhibiting elastase that has already been released and preventing the breakdown of TGF- β ³¹. Furthermore, SLPI binds and synergizes with proepithelin, a cytokine that promotes epithelial growth and suppresses neutrophil activation, protecting it from

proteolytic conversion into pro-inflammatory epithelins³⁶. CD44-positive macrophages clear the hyaluronan fragments. Fresh neutrophils no longer enter the site, and those present undergo apoptosis. Macrophages ingest apoptotic neutrophils and degrade their residual stores of elastase. TNF induces macrophages to release interleukin-12, which induces lymphocytes to release interferon- γ (IFN- γ). IFN- γ acts early on to induce macrophage chemokine production, but now suppresses it³³. Ingestion of apoptotic neutrophils elicits more TGF- β from macrophages, and the predominant action of TGF- β is no longer the recruitment of neutrophils, but instead the promotion of tissue repair. Thus, TNF, IFN- γ and TGF- β join PGE2 as examples of molecules whose actions switch from pro-inflammatory to anti-inflammatory, depending on timing and context.

ROIs and RNIs are two additional sets of molecules that can either promote or suppress inflammation. Chronic granulomatous disease (CGD), a genetic disorder predisposing to life-threatening bacterial and fungal infections, results from a deficiency in the ROI-producing enzyme phagocyte oxidase (phox). In CGD, chronic inflammation sometimes seems to precede infection or long outlast it³⁷. The clinical impression of an exaggerated granulomatous response in CGD has been confirmed in phox-deficient mice, which form abnormally large granulomas when injected with sterile fungal cell walls³⁸. These observations demonstrate that phox has an important anti-inflammatory role, such as oxidatively inactivating chemotactic factors³⁹, even though phox can be profoundly pro-inflammatory by virtue of oxidizing tissue constituents, oxidatively activating metalloproteinases and oxidatively inactivating protease inhibitors¹⁷. Similarly, mice deficient in inducible nitric oxide synthase (iNOS) display a triple phenotype — increased susceptibility to infection, reduced inflammation or excessive inflammation — depending on the experimental setting^{40,41}. Without infection or other experimental intervention, however, mice lacking phox or iNOS appear normal, in contrast to the situation discussed next.

Genes whose disruption predisposes to inflammation

Another level of control is revealed by the fact that there are numerous genes whose disruption predisposes to inflammation in people or mice living under conventional conditions without evident provocations that are known to elicit inflammation in wild-type hosts (Table 2).

That loss-of-function mutations can lead to spontaneous inflammation was probably shown first and most clearly by human C1q deficiency⁴². This disorder confers a >90% incidence of systemic lupus erythematosus⁴³. That Table 2, although incomplete, includes over 50 genes implies that health does not arise passively from a lack of pro-inflammatory stimuli. On the contrary, potentially inflammatory stimuli seem to be ubiquitous, and it takes an active process to avoid over-reacting to those that pose a minimal threat.

The diverse genes necessary to suppress spontaneous inflammation can be classified into functional sets. Although such groupings are subjective, it is difficult to posit less than three key elements of the tonic anti-inflammatory state. First is the solubilization and clearance of immune complexes and cellular debris. The second element is the balanced progression of leukocytes and lymphocytes through programmes of activation, proliferation and apoptosis. Third is the avoidance of oxidative injury, such as by disposal of haem and by constitutive restraint on the respiratory burst activity of macrophages that are continually exposed to particulate stimuli⁴⁴.

Gene products are included in Table 2 because of their demonstrated importance for avoiding inflammation. In a paradox that is now familiar, most of these proteins, such as TNF-R1 and NF- κ B, are better known for their essential contributions to promoting inflammation. Again it emerges that pro-inflammatory gene products are often essential effectors of anti-inflammatory homeostasis.

Many of the phenotypes listed in Table 2 are strongly dependent on genetic background, age, sex and environmental conditions (such as intestinal flora). The profound influence of epistatic or nongenetic

factors is apparent when considering the following contrasts. First, mice of one strain contrasted with mice of another, such as those whose deficiency in interleukin-1 receptor antagonist (IL-1Ra) leads either to arthritis or arteritis. Second, one person contrasted with another, such as those with NOD2 mutations who develop either enterocolitis or arthritis (NOD2 is an intracellular lipopolysaccharide- and kinase-binding protein with a caspase-recruitment domain (CARD), a nucleotide-binding domain and leucine-rich repeats (LRRs)⁴⁵). Third, people contrasted with mice, such as those whose disrupted C4 gene predisposes to rash or glomerulonephritis, respectively.

It is a puzzle that disrupting a given gene has tissue-specific consequences when expression of the gene in question is not restricted to, or in some cases even manifest by, the tissue that is inflamed. This is not fully explained by the ability of leukocytes and lymphocytes to enter any tissue. Overall, the sites most frequently affected by inflammation in association with the listed mutations (skin~lung > kidney > joints > colon~liver > heart > pancreas ~ eyes > other organs) are those that are anatomically large (skin, lung, colon and liver), continually exposed to microbes (skin, lung, colon and conjunctivae), or prone to trapping immune complexes (kidney, joints and skin).

Finally, Table 2 can be viewed as a collection of mechanistic mysteries that invite investigation. Among the most intriguing are the genes whose mutation predisposes to periodic fever syndromes. Cryopyrin resembles NOD2, Toll-like receptors and CD14 in carrying an LRR. Pyrin and cryopyrin contain pyrin domains, which are predicted to share structural features with CARDs, such as those present in NOD2⁴⁶. What is the mechanism of the anti-inflammatory actions of these proteins? Do they have pro-inflammatory actions as well? What are the contributions of their LRR, pyrin and CARD domains? Why do the phenotypes of their mutations mimic those associated with mutations in genes encoding two additional and very different proteins, TNF-R1 and mevalonate kinase? Mevalonate kinase is essential for synthesis of isoprenoids and cholesterol. Perhaps this sheds light on the unexplained anti-inflammatory effects of statins, drugs that block cholesterol synthesis at a later step⁴⁷. Administration of statins might lead to accumulation of a cholesterol precursor whose formation depends on mevalonate kinase. Perhaps this intermediate has potent anti-inflammatory actions on its own, or confers such actions on a protein to which it becomes attached.

Perspective

Our bodies sustain the replication of hundreds of different genomes. Only the largest is heritable in the germ line. To preserve its opportunity to be transmitted, the germ line genome must encode hair-trigger vigilance against take-over of the soma by genomes that replicate far faster. The rapid mobilization of microbicidal defences evolved at the cost of potentially suicidal autotoxicity (see review in this issue by Cohen, pages 885–891). Thus, for host survival, two sets of mechanisms must be matched: the ability to mount a rapid inflammatory response to injurious microbial invasion, and the ability to refrain from doing so otherwise.

For those seeking the origins of inflammatory or autoimmune diseases, this analysis encourages two lines of inquiry: First, what might predispose to the formation, modification or relocalization of endogenous molecules such that they activate detection systems that normally report injury and infection? Second, are there dysfunctions in pathways whose integrity is required to prevent inflammation from arising spontaneously?

For those trying to promote inflammation, the analysis offered above commends combinations of signals that mimic both injury of the host and the presence of infectious agents, without resorting to either.

For those developing anti-inflammatory therapies, the need for each of several go signals suggests that it should be relatively straightforward to interrupt inflammation. Unfortunately, the redundancy of many signals (over-determination) complicates this goal. The

recognition of stop signals offers additional opportunities to abort inflammation⁴⁸. However, predictability is complicated by the tendency of signals to shift sense, as illustrated for PGE2, TNF, IFN- γ , TGF- β , ROIs and RNIs. Finally, there remains the dilemma that the more broadly an agent suppresses inflammation, the more likely it will exacerbate infections. Corticosteroids taught this lesson, and TNF-neutralizing agents reinforced it⁴⁹. Nonetheless, many of those working in anti-inflammatory research are optimistic. Experimental biology is uncovering an unprecedented wealth of molecular detail at a time when systems biology seems poised to put into perspective the complexity and dynamics of the inflammatory process. An alliance between experimental and systems biology should be a powerful force to identify points of control amenable to relatively safe and effective intervention. □

doi:10.1038/nature01320

- Zweifach, B. W., Grant, L. & McCluskey, R. T. *The Inflammatory Process* (Academic, New York, 1965).
- Bunting, M., Harris, E. S., McIntyre, T. M., Prescott, S. M. & Zimmerman, G. A. Leukocyte adhesion deficiency syndromes: adhesion and tethering defects involving β_2 integrins and selectin ligands. *Curr. Opin. Hematol.* **9**, 30–35 (2002).
- Biesma, D. H. et al. A family with complement factor D deficiency. *J. Clin. Invest.* **108**, 233–240 (2001).
- Soiffer, R. et al. Vaccination with irradiated autologous melanoma cells engineered to secrete human granulocyte-macrophage colony-stimulating factor generates potent antitumor immunity in patients with metastatic melanoma. *Proc. Natl Acad. Sci. USA* **95**, 13141–13146 (1998).
- Morales, A. Intravesical therapy of bladder cancer: an immunotherapy success story. *Int. J. Urol.* **3**, 329–333 (1996).
- Riewald, M., Petrovan, R. J., Donner, A., Mueller, B. M. & Ruf, W. Activation of endothelial cell protease activated receptor 1 by the protein C pathway. *Science* **296**, 1880–1882 (2002).
- Steinhoff, M. et al. Agonists of proteinase-activated receptor 2 induce inflammation by a neurogenic mechanism. *Nature Med.* **6**, 151–158 (2000).
- Basu, S. & Srivastava, P. K. Heat shock proteins: the fountainhead of innate and adaptive immune responses. *Cell Stress Chaper.* **5**, 443–451 (2000).
- Scaffidi, P., Misteli, T. & Bianchi, M. E. Release of chromatin protein HMGB1 by necrotic cells triggers inflammation. *Nature* **418**, 191–195 (2002).
- Carp, H. Mitochondrial N-formylmethionyl proteins as chemoattractants for neutrophils. *J. Exp. Med.* **155**, 264–275 (1982).
- Muller, W. A. Leukocyte-endothelial cell interactions in the inflammatory response. *Lab. Invest.* **82**, 521–533 (2002).
- Lee, D. M. et al. Mast cells: a cellular link between autoantibodies and inflammatory arthritis. *Science* **297**, 1689–1692 (2002).
- van der Poll, T. Coagulation and inflammation. *J. Endotoxin Res.* **7**, 301–304 (2001).
- Kaplan, A. P., Joseph, K. & Silverberg, M. Pathways for bradykinin formation and inflammatory disease. *J. Allergy Clin. Immunol.* **109**, 195–209 (2002).
- Nathan, C., Xie, Q. W., Halbwachs-Mecarelli, L. & Jin, W. W. Albumin inhibits neutrophil spreading and hydrogen peroxide release by blocking the shedding of CD43 (sialoporphin, leukosialin). *J. Cell Biol.* **122**, 243–256 (1993).
- Nathan, C. F. Neutrophil activation on biological surfaces. Massive secretion of hydrogen peroxide in response to products of macrophages and lymphocytes. *J. Clin. Invest.* **80**, 1550–1560 (1987).
- Weiss, S. J. Tissue destruction by neutrophils. *N. Engl. J. Med.* **320**, 365–376 (1989).
- Morgan, J. G., Pereira, H. A., Sukiennicki, T., Spitznagel, J. K. & Larrick, J. W. Human neutrophil granule cationic protein CAP37 is a specific macrophage chemotaxin that shares homology with inflammatory proteinases. *Adv. Exp. Med. Biol.* **305**, 89–96 (1991).
- Yang, D. et al. β -defensins: linking innate and adaptive immunity through dendritic and T cell CCR6. *Science* **286**, 525–528 (1999).
- Robbiani, D. F. et al. The leukotriene C_4 transporter MRP1 regulates CCL19 (MIP-3 β , ELC)-dependent mobilization of dendritic cells to lymph nodes. *Cell* **103**, 757–768 (2000).
- Matzinger, P. The danger model: a renewed sense of self. *Science* **296**, 301–305 (2002).
- Medzhitov, R. & Janeway, C. A. Jr Decoding the patterns of self and nonself by the innate immune system. *Science* **296**, 298–300 (2002).
- Kiehl, S. et al. Toll-like receptor 4 polymorphisms and atherogenesis. *N. Engl. J. Med.* **347**, 185–192 (2002).
- Fink, M. P. Effect of critical illness on microbial translocation and gastrointestinal mucosa permeability. *Semin. Respir. Infect.* **9**, 256–260 (1994).
- Levy, B. D., Clish, C. B., Schmidt, B., Gronert, K. & Serhan, C. N. Lipid mediator class switching during acute inflammation: signals in resolution. *Nature Immunol.* **2**, 612–619 (2001).
- Mizumoto, N. et al. CD39 is the dominant Langerhans cell-associated ecto-ATPase: modulatory roles in inflammation and immune responsiveness. *Nature Med.* **8**, 358–365 (2002).
- Ohta, A. & Sitkovsky, M. Role of G-protein-coupled adenosine receptors in downregulation of inflammation and protection from tissue damage. *Nature* **414**, 916–920 (2001).
- Hoek, R. M. et al. Down-regulation of the macrophage lineage through interaction with OX2 (CD200). *Science* **290**, 1768–1771 (2000).
- Daheshia, M., Friend, D. S., Grusby, M. J., Austen, K. F. & Katz, H. R. Increased severity of local and systemic anaphylactic reactions in gp49B1-deficient mice. *J. Exp. Med.* **194**, 227–234 (2001).
- Teder, P. et al. Resolution of lung inflammation by CD44. *Science* **296**, 155–158 (2002).
- Ashcroft, G. S. et al. Secretory leukocyte protease inhibitor mediates non-redundant functions necessary for normal wound healing. *Nature Med.* **6**, 1147–1153 (2000).
- Marino, M. W. et al. Characterization of tumor necrosis factor-deficient mice. *Proc. Natl Acad. Sci. USA* **94**, 8093–8098 (1997).
- Hodge-Dufour, J. et al. Inhibition of interferon γ induced interleukin 12 production: a potential mechanism for the anti-inflammatory activities of tumor necrosis factor. *Proc. Natl Acad. Sci. USA* **95**, 13806–13811 (1998).

34. Jin, F. Y., Nathan, C., Radzioch, D. & Ding, A. Secretory leukocyte protease inhibitor: a macrophage product induced by and antagonistic to bacterial lipopolysaccharide. *Cell* **88**, 417–426 (1997).
35. Grobmyer, S. R. *et al.* Secretory leukocyte protease inhibitor, an inhibitor of neutrophil activation, is elevated in serum in human sepsis and experimental endotoxemia. *Crit. Care Med.* **28**, 1276–1282 (2000).
36. Zhu, J. *et al.* Conversion of proepithelin to epithelins: roles of SLP1 and elastase in host defense and wound repair. *Cell* (in the press).
37. Segal, B. H., Leto, T. L., Gallin, J. I., Malech, H. L. & Holland, S. M. Genetic, biochemical, and clinical features of chronic granulomatous disease. *Medicine (Baltimore)* **79**, 170–200 (2000).
38. Morgenstern, D. E., Gifford, M. A., Li, L. L., Doerschuk, C. M. & Dinamer, M. C. Absence of respiratory burst in X-linked chronic granulomatous disease mice leads to abnormalities in both host defense and inflammatory response to *Aspergillus fumigatus*. *J. Exp. Med.* **185**, 207–218 (1997).
39. Clark, R. A. & Klebanoff, S. J. Chemotactic factor inactivation by the myeloperoxidase-hydrogen peroxide-halide system. *J. Clin. Invest.* **64**, 913–920 (1979).
40. Nathan, C. Inducible nitric oxide synthase: what difference does it make? *J. Clin. Invest.* **100**, 2417–2423 (1997).
41. Bogdan, C. Nitric oxide and the immune response. *Nature Immunol.* **2**, 907–916 (2001).
42. Schur, P. H. Genetics of complement deficiencies associated with lupus-like syndromes. *Arthritis Rheum.* **21**, S153–S160 (1978).
43. Walport, M. J., Davies, K. A., Morley, B. J. & Botto, M. Complement deficiency and autoimmunity. *Ann. NY Acad. Sci.* **815**, 267–281 (1997).
44. Wert, S. E. *et al.* Increased metalloproteinase activity, oxidant production, and emphysema in surfactant protein D gene-inactivated mice. *Proc. Natl Acad. Sci. USA* **97**, 5972–5977 (2000).
45. Ogura, Y. *et al.* Nod2, a Nod1/Apaf-1 family member that is restricted to monocytes and activates NF- κ B. *J. Biol. Chem.* **276**, 4812–4818 (2001).
46. Kastner, D. L. & O'Shea, J. A fever gene comes in from the cold. *Nature Genet.* **29**, 241–242 (2001).
47. Liao, J. K. Isoprenoids as mediators of the biological effects of statins. *J. Clin. Invest.* **110**, 285–288 (2002).
48. Levy, B. D. *et al.* Multi-pronged inhibition of airway hyper-responsiveness and inflammation by lipoxin A₂. *Nature Med.* **8**, 1018–1023 (2002).
49. Keane, J. *et al.* Tuberculosis associated with infliximab, a tumor necrosis factor α -neutralizing agent. *N. Engl. J. Med.* **345**, 1098–1104 (2001).
50. Vaishnaw, A. K. *et al.* The spectrum of apoptotic defects and clinical manifestations, including systemic lupus erythematosus, in humans with CD95 (Fas/APO-1) mutations. *Arthritis Rheum.* **42**, 1833–1842 (1999).
51. Fisher, G. H. *et al.* Dominant interfering Fas gene mutations impair apoptosis in a human autoimmune lymphoproliferative syndrome. *Cell* **81**, 935–946 (1995).
52. Sneller, M. C. *et al.* A novel lymphoproliferative/autoimmune syndrome resembling murine lpr/gld disease. *J. Clin. Invest.* **90**, 334–341 (1992).
53. Cohen, P. L. & Eisenberg, R. A. lpr and gld: single gene models of systemic autoimmunity and lymphoproliferative disease. *Annu. Rev. Immunol.* **9**, 243–269 (1991).
54. Botto, M. *et al.* Homozygous C1q deficiency causes glomerulonephritis associated with multiple apoptotic bodies. *Nature Genet.* **19**, 56–59 (1998).
55. Lipsker, D. M. *et al.* Lupus erythematosus associated with genetically determined deficiency of the second component of the complement. *Arch. Dermatol.* **136**, 1508–1514 (2000).
56. Chen, Z., Koralov, S. B. & Kelsoe, G. Complement C4 inhibits systemic autoimmunity through a mechanism independent of complement receptors CR1 and CR2. *J. Exp. Med.* **192**, 1339–1352 (2000).
57. Sullivan, K. E. Complement deficiency and autoimmunity. *Curr. Opin. Pediatr.* **10**, 600–606 (1998).
58. Xu, C. *et al.* A critical role for murine complement regulator Crry in fetomaternal tolerance. *Science* **287**, 498–501 (2000).
59. Bickerstaff, M. C. *et al.* Serum amyloid P component controls chromatin degradation and prevents antinuclear autoimmunity. *Nature Med.* **5**, 694–697 (1999).
60. Napirei, M. *et al.* Features of systemic lupus erythematosus in Dnase1-deficient mice. *Nature Genet.* **25**, 177–181 (2000).
61. Bolland, S., Yim, Y. S., Tus, K., Wakeland, E. K. & Ravetch, J. V. Genetic modifiers of systemic lupus erythematosus in Fc γ RIIB^{-/-} mice. *J. Exp. Med.* **195**, 1167–1174 (2002).
62. Sullivan, K. E., Mullen, C. A., Blaes, R. M. & Winkelstein, J. A. A multi-institutional survey of the Wiskott-Aldrich syndrome. *J. Pediatr.* **125**, 876–885 (1994).
63. Leverier, Y. *et al.* Cutting edge: the Wiskott-Aldrich syndrome protein is required for efficient phagocytosis of apoptotic cells. *J. Immunol.* **166**, 4831–4834 (2001).
64. Snapper, S. B. *et al.* Wiskott-Aldrich syndrome protein-deficient mice reveal a role for WASP in T but not B cell activation. *Immunity* **9**, 81–91 (1998).
65. McDermott, M. F. *et al.* Germline mutations in the extracellular domains of the 55 kDa TNF receptor, TNFR1, define a family of dominantly inherited autoinflammatory syndromes. *Cell* **97**, 133–144 (1999).
66. Shull, M. M. *et al.* Targeted disruption of the mouse transforming growth factor- β 1 gene results in multifocal inflammatory disease. *Nature* **359**, 693–699 (1992).
67. Kulkarni, A. B. *et al.* Transforming growth factor β , null mutation in mice causes excessive inflammatory response and early death. *Proc. Natl Acad. Sci. USA* **90**, 770–774 (1993).
68. Willerford, D. M. *et al.* Interleukin-2 receptor α chain regulates the size and content of the peripheral lymphoid compartment. *Immunity* **3**, 521–530 (1995).
69. Sadlack, B. *et al.* Ulcerative colitis-like disease in mice with a disrupted interleukin-2 gene. *Cell* **75**, 253–261 (1993).
70. Kuhn, R., Lohler, J., Rennick, D., Rajewsky, K. & Muller, W. Interleukin-10-deficient mice develop chronic enterocolitis. *Cell* **75**, 263–274 (1993).
71. Dranoff, G. *et al.* Involvement of granulocyte-macrophage colony-stimulating factor in pulmonary homeostasis. *Science* **264**, 713–716 (1994).
72. Nicklin, M. J., Hughes, D. E., Barton, J. L., Ure, J. M. & Duff, G. W. Arterial inflammation in mice lacking the interleukin 1 receptor antagonist gene. *J. Exp. Med.* **191**, 303–312 (2000).
73. Horai, R. *et al.* Development of chronic inflammatory arthropathy resembling rheumatoid arthritis in interleukin 1 receptor antagonist-deficient mice. *J. Exp. Med.* **191**, 313–320 (2000).
74. Mombaerts, P. *et al.* Spontaneous development of inflammatory bowel disease in T cell receptor mutant mice. *Cell* **75**, 274–282 (1993).
75. Waterhouse, P. *et al.* Lymphoproliferative disorders with early lethality in mice deficient in *Ctla-4*. *Science* **270**, 985–988 (1995).
76. Tivol, E. A. *et al.* Loss of CTLA-4 leads to massive lymphoproliferation and fatal multiorgan tissue destruction, revealing a critical negative regulatory role of CTLA-4. *Immunity* **3**, 541–547 (1995).
77. Nishimura, H., Nose, M., Hiai, H., Minato, N. & Honjo, T. Development of lupus-like autoimmune diseases by disruption of the PD-1 gene encoding an ITIM motif-carrying immunoreceptor. *Immunity* **11**, 141–151 (1999).
78. Sommers, C. L. *et al.* A LAT mutation that inhibits T cell development yet induces lymphoproliferation. *Science* **296**, 2040–2043 (2002).
79. Alexander, W. S. *et al.* SOCS1 is a critical inhibitor of interferon γ signaling and prevents the potentially fatal neonatal actions of this cytokine. *Cell* **98**, 597–608 (1999).
80. Okkenhaug, K. *et al.* Impaired B and T cell antigen receptor signaling in p110 δ PI 3-kinase mutant mice. *Science* **297**, 1031–1034 (2002).
81. Di Cristofano, A. *et al.* Impaired Fas response and autoimmunity in *Pten*^{+/-} mice. *Science* **285**, 2122–2125 (1999).
82. Hibbs, M. L. *et al.* Multiple defects in the immune system of Lyn-deficient mice, culminating in autoimmune disease. *Cell* **83**, 301–311 (1995).
83. Nishizumi, H. *et al.* Impaired proliferation of peripheral B cells and indication of autoimmune disease in lyn-deficient mice. *Immunity* **3**, 549–560 (1995).
84. Bachmaier, K. *et al.* Negative regulation of lymphocyte activation and autoimmunity by the molecular adaptor Cbl-b. *Nature* **403**, 211–216 (2000).
85. Rudolph, U. *et al.* Ulcerative colitis and adenocarcinoma of the colon in Γ α ₂-deficient mice. *Nature Genet.* **10**, 143–150 (1995).
86. Tsui, H. W., Siminovich, K. A., de Souza, L. & Tsui, F. W. *Motheaten* and *viable motheaten* mice have mutations in the haematopoietic cell phosphatase gene. *Nature Genet.* **4**, 124–129 (1993).
87. Shultz, L. D. *et al.* Mutations at the murine motheaten locus are within the hematopoietic cell protein-tyrosine phosphatase (Hcph) gene. *Cell* **73**, 1445–1454 (1993).
88. Helgason, C. D. *et al.* Targeted disruption of SHIP leads to hemopoietic perturbations, lung pathology, and a shortened life span. *Genes Dev.* **12**, 1610–1620 (1998).
89. Balomenos, D. *et al.* The cell cycle inhibitor p21 controls T-cell proliferation and sex-linked lupus development. *Nature Med.* **6**, 171–176 (2000).
90. Taylor, G. A. *et al.* A pathogenic role for TNF α in the syndrome of cachexia, arthritis, and autoimmunity resulting from tristetraprolin (TTP) deficiency. *Immunity* **4**, 445–454 (1996).
91. Ranger, A. M., Oukka, M., Rengarajan, J. & Glimcher, L. H. Inhibitory function of two NFAT family members in lymphoid homeostasis and Th2 development. *Immunity* **9**, 627–635 (1998).
92. Pasparrakis, M. *et al.* TNF-mediated inflammatory skin disease in mice with epidermis-specific deletion of IKK2. *Nature* **417**, 861–866 (2002).
93. The International Incontinentia Pigmenti (IP) Consortium. Genomic rearrangement in *NEMO* impairs NF- κ B activation and is a cause of incontinentia pigmenti. *Nature* **405**, 466–472 (2000).
94. Schmidt-Suppran, M. *et al.* NEMO/IKK γ -deficient mice model incontinentia pigmenti. *Mol. Cell* **5**, 981–992 (2000).
95. Makris, C. *et al.* Female mice heterozygous for IKK γ /NEMO deficiencies develop a dermatopathy similar to the human X-linked disorder incontinentia pigmenti. *Mol. Cell* **5**, 969–979 (2000).
96. Klement, J. F. *et al.* IkB α deficiency results in a sustained NF- κ B response and severe widespread dermatitis in mice. *Mol. Cell Biol.* **16**, 2341–2349 (1996).
97. Perry, W. L. *et al.* The itchy locus encodes a novel ubiquitin protein ligase that is disrupted in a18H mice. *Nature Genet.* **18**, 143–146 (1998).
98. Weih, F. *et al.* Multiorgan inflammation and hematopoietic abnormalities in mice with a targeted disruption of RelB, a member of the NF- κ B/Rel family. *Cell* **80**, 331–340 (1995).
99. Burkly, L. *et al.* Expression of *relB* is required for the development of thymic medulla and dendritic cells. *Nature* **373**, 531–536 (1995).
100. Barton, D., HogenEsch, H. & Weih, F. Mice lacking the transcription factor RelB develop T cell-dependent skin lesions similar to human atopic dermatitis. *Eur. J. Immunol.* **30**, 2323–2332 (2000).
101. Ishikawa, H. *et al.* Chronic inflammation and susceptibility to bacterial infections in mice lacking the polypeptide (p)105 precursor (NF- κ B1) but expressing p50. *J. Exp. Med.* **187**, 985–996 (1998).
102. Finotto, S. *et al.* Development of spontaneous airway changes consistent with human asthma in mice lacking T-bet. *Science* **295**, 336–338 (2002).
103. Salvador, J. M. *et al.* Mice lacking the p53-effector gene *Gadd45a* develop a lupus-like syndrome. *Immunity* **16**, 499–508 (2002).
104. Panwala, C. M., Jones, J. C. & Viney, J. L. A novel model of inflammatory bowel disease: mice deficient for the multiple drug resistance gene, *mdr1a*, spontaneously develop colitis. *J. Immunol.* **161**, 5733–5744 (1998).
105. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
106. Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
107. Miceli-Richard, C. *et al.* CARD15 mutations in Blau syndrome. *Nature Genet.* **29**, 19–20 (2001).
108. International Mediterranean Fever Consortium. Ancient missense mutations in a new member of the *RoRet* gene family are likely to cause familial Mediterranean fever. *Cell* **90**, 797–807 (1997).
109. French FMF Consortium. A candidate gene for familial Mediterranean fever. *Nature Genet.* **17**, 25–31 (1997).
110. Hoffman, H. M., Mueller, J. L., Broide, D. H., Wanderer, A. A. & Kolodner, R. D. Mutation of a new gene encoding a putative pyrin-like protein causes familial cold autoinflammatory syndrome and Muckle-Wells syndrome. *Nature Genet.* **29**, 301–305 (2001).
111. Houten, S. M. *et al.* Mutations in *MVK*, encoding mevalonate kinase, cause hyperimmunoglobulinemia D and periodic fever syndrome. *Nature Genet.* **22**, 175–177 (1999).
112. Drenth, J. P. *et al.* Mutations in the gene encoding mevalonate kinase cause hyper-IgD and periodic fever syndrome. International Hyper-IgD Study Group. *Nature Genet.* **22**, 178–181 (1999).
113. Poss, K. D. & Tonegawa, S. Heme oxygenase 1 is required for mammalian iron reutilization. *Proc. Natl Acad. Sci. USA* **94**, 10919–10924 (1997).

Acknowledgements

I thank L. Grant for introducing me to the study of inflammation, K. F. Austen, P. Bernstein, A. Ding, M. Fuortes and L. Old for critique of the paper and S. Chen for help in the library. It is regretted that space precluded citing many relevant sources. Preparation of this article was supported by NIH. The Department of Microbiology and Immunology acknowledges the support of the William Randolph Hearst Foundation.

The inflammatory reflex

Kevin J. Tracey

Laboratory of Biomedical Science, North Shore-LIJ Research Institute, 350 Community Drive, Manhasset, New York 11030, USA

(e-mail: kjtracey@sprynet.com)

Inflammation is a local, protective response to microbial invasion or injury. It must be fine-tuned and regulated precisely, because deficiencies or excesses of the inflammatory response cause morbidity and shorten lifespan. The discovery that cholinergic neurons inhibit acute inflammation has qualitatively expanded our understanding of how the nervous system modulates immune responses. The nervous system reflexively regulates the inflammatory response in real time, just as it controls heart rate and other vital functions. The opportunity now exists to apply this insight to the treatment of inflammation through selective and reversible 'hard-wired' neural systems.

"The mind has great influence over the body, and maladies often have their origin there." Molière (1622–1673).

Survival is impossible without vigilant defence against attack and injury. The innate immune system continuously surveys the body for the presence of invaders. When it encounters an attack, it involuntarily sets in motion a discrete, localized inflammatory response to thwart most pathogenic threats. The magnitude of the inflammatory response is crucial: insufficient responses result in immunodeficiency, which can lead to infection and cancer; excessive responses cause morbidity and mortality in diseases such as rheumatoid arthritis, Crohn's disease, atherosclerosis, diabetes, Alzheimer's disease, multiple sclerosis, and cerebral and myocardial ischaemia. If inflammation spreads into the bloodstream, as occurs in septic shock syndrome, sepsis, meningitis and severe trauma, the inflammatory responses can be more dangerous than the original inciting stimulus. Homeostasis and health are restored when inflammation is limited by anti-inflammatory responses that are redundant, rapid, reversible, localized, adaptive to changes in input and integrated by the nervous system.

The nervous system is composed of sensory systems (which detect the state of the body and organs) and motor systems (which transmit signals to the body and organs). Whereas the somatic motor system controls voluntary movements, the autonomic motor system controls visceral body functions and innervates glands (involuntary). The autonomic nervous system has two principal divisions, the parasympathetic pathway and the sympathetic pathway, which act either in synergy or in opposition to mediate basic physiological responses in real time. The autonomic system continuously controls heart rate and blood pressure, respiratory rate, gastrointestinal motility, body temperature and other constantly changing, essential life functions. The autonomic nervous system interacts with the primitive brain, including the limbic system (serving important memory functions), brain stem and hypothalamus. Hypothalamic neural output is relayed to sympathetic and parasympathetic nuclei in the brain stem and spinal cord. Hormonal input also controls the release of pituitary hormones, which in turn regulate basic functions of the endocrine organs. Autonomic nervous functions are normally subconscious, but essential basic autonomic functions can be placed under conscious control from signals originating in the higher brain (cerebral cortex). For example, subjects can be trained through biofeedback to

decrease their heart rate by increasing parasympathetic outflow.

Recent insights have identified a basic neural pathway that reflexively monitors and adjusts the inflammatory response. Inflammatory stimuli activate sensory pathways that relay information to the hypothalamus. Like a knee-jerk reflex, in which the stretching of a patellar tendon elicits a rapidly opposing motor action, inflammatory input activates an anti-inflammatory response that is fast and subconscious. This prevents spillage of inflammatory products into the circulation. The nervous system integrates the inflammatory response: it gathers information about invasive events from several local sites, mobilizes defences and creates memory to improve chances for survival.

Here I review evidence showing that the neural control of acute inflammation is reflexive, directly interconnected and controllable. Special emphasis is placed on cholinergic anti-inflammatory mechanisms that inhibit the activation of macrophages and the release of cytokines (Fig. 1). I also discuss evidence indicating that stimulation of the vagus nerve, by either electrical or pharmacological means, prevents inflammation and inhibits the release of cytokines that are clinically relevant drug targets for treating inflammatory disease.

Inflammation mediated by TNF

Tumour-necrosis factor (TNF), a cytokine with a relative molecular mass of 17,000 (M_r 17K), is produced by activated macrophages in response to pathogens and other injurious stimuli, and is a necessary and sufficient mediator of local and systemic inflammation^{1,2}. Local increases in TNF cause the cardinal clinical signs of inflammation, including heat, swelling, pain and redness. Systemic increases in TNF mediate tissue injury by depressing cardiac output, inducing microvascular thrombosis and mediating systemic capillary leakage syndrome. TNF amplifies and prolongs the inflammatory response by activating other cells to release both cytokines such as interleukin 1 (IL-1) and high mobility group B1 (HMGB1), and mediators such as eicosanoids, nitric oxide and reactive oxygen species, which promote further inflammation and tissue injury³. TNF is essential for the complete expression of inflammation during invasion, and self-limited inflammation is normally characterized by decreasing TNF activity.

Low amounts of TNF can contribute to host defence by limiting the spread of pathogenic organisms into the

Table 1 'Diffusible' anti-inflammatory mediators

Cytokines	IL-10 TGF- β TNF-binding protein IL-1R α
Hormones	Glucocorticoids Adrenaline Noradrenaline α -MSH
Local effectors	Spermine Prostaglandin E2 Fetuin Heat-shock proteins Acute phase proteins

circulation, promoting coagulation to localize the invader, and stimulating the growth of damaged tissues⁴. In a typical 'successful' inflammatory response, the duration and magnitude of TNF release is limited, its beneficial and protective activities predominate, and it is not released systemically. Studies of the inflammatory action of TNF in non-malignant disease have led to widespread investigation of both the 'normal' mechanisms that regulate inflammation and the therapeutic potential of monoclonal antibodies specific for TNF.

Monoclonal antibodies against TNF

Early studies using monoclonal antibodies against TNF showed that this approach effectively prevents lethal tissue injury during bacterial invasion¹. Subsequent clinical trials led to the registration of both monoclonal antibodies against TNF, and TNF-binding proteins for treating rheumatoid arthritis and Crohn's disease. Many individuals with these debilitating inflammatory illnesses have enjoyed disease remissions and an improved quality of life. Crippling joint pain has been alleviated in children with rheumatoid arthritis treated with TNF antibodies; some of the youngest patients have even experienced 'catch-up growth' and normalization of development (U. Andersson, personal communication). These and other clinical successes with TNF monoclonal antibodies have proved that cytokine responses can be manipulated to specific therapeutic advantage for inflammatory disease.

But strategies using TNF antibodies have not been translated successfully into treatments for bacterial invasion or sepsis, for reasons that have been reviewed extensively elsewhere⁵. Most notably, serum concentrations of TNF were undetectable in most of the individuals recruited into clinical sepsis trials, because the study group comprised a heterogeneous population with diverse diseases at varying stages of illness. In early experiments of the use of TNF monoclonal antibodies in bacteraemia, it became clear that TNF is an early mediator of inflammation and that TNF antibodies are ineffective if therapy is initiated after serum TNF has been cleared¹. Continued interest in understanding the use of TNF antibodies for individuals with sepsis is now focused on identifying a homogenous study population with increased serum TNF for treatment early in the course of illness.

An alternative therapeutic strategy is to target 'late' mediators of lethality that are produced after TNF in the inflammatory pathway³. HMGB1 has been implicated as an experimental therapeutic target that is produced relatively late in the course of endotoxaemia^{3,6}. Antibodies specific for HMGB1 confer significant protection against the lethality of endotoxaemia, even when the first antibody doses are administered after the early TNF concentrations have been cleared. Reducing serum concentrations of HMGB1 by administering ethyl pyruvate as late as 24 h after the onset of sepsis rescues animals from lethality, indicating that it may now be possible to develop therapies for sepsis that cover a significantly wider, clinically relevant treatment window^{7,8}.

Other cytokines have been implicated as therapeutic targets in the pathogenesis of inflammatory diseases, and it is likely that future treatment plans will target mediators in addition to TNF. Although I focus the discussion of neural regulation of inflammation primarily on cholinergic inhibition of TNF, evidence indicates that these neural

anti-inflammatory mechanisms also inhibit the release of IL-1, IL-18 and HMGB1.

Anti-inflammatory responses normally inhibit inflammation

Highly conserved, counter-regulatory mechanisms normally limit the acute inflammatory response and prevent the spread of inflammatory mediators into the bloodstream (Table 1). Activated immunologically competent cells release TNF receptor fragments that bind and neutralize its inflammatory and potentially toxic actions⁹. Anti-inflammatory cytokines, such as IL-10 and transforming growth factor- β (TGF- β), specifically inhibit the release of TNF and other proinflammatory mediators¹⁰. Adrenal glucocorticoids, adrenaline, α -melanocyte-stimulating hormone (α -MSH) and other 'classical' stress hormones inhibit cytokine synthesis and intracellular signal transduction¹¹⁻¹⁴. Spermine accumulates at sites of tissue injury and infection and inhibits macrophage activation and cytokine synthesis¹⁵.

The importance of these endogenous anti-inflammatory pathways has been shown by experimentally impairing isolated pathways. For example, animals subjected to hypophysectomy or adrenalectomy are significantly sensitized to the lethal effects of endotoxin¹⁶. In the absence of an adequate adrenocorticotrophic hormone (ACTH) and glucocorticoid response, TNF is significantly overexpressed during endotoxaemia¹⁶⁻¹⁸. Functional deficiencies in the release of corticotropin-releasing factor (CRF) predispose Lewis rats to developing experimental arthritis induced by streptococcal antigens because of an insufficient glucocorticoid response^{13,19,20}. Animals deficient in IL-10 develop a chronic inflammatory bowel disease that predominately affects the colon²¹ and are susceptible to a more severe form of collagen-induced arthritis²². Administration of specific pharmacological spermine antagonists significantly increases local TNF activity and carrageenan-induced oedema formation, and amplifies the inflammatory response¹⁵. Together, these findings show that loss of endogenous anti-inflammatory mechanisms converts a normally protective, self-limited inflammatory response into an excessive, potentially deleterious response.

Communication between immune and nervous systems

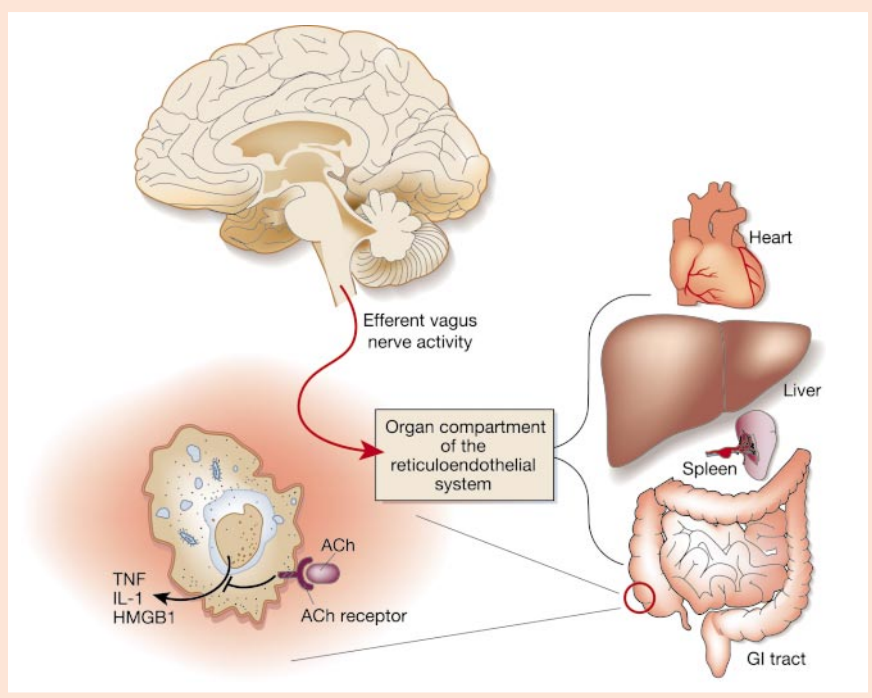
The activation of pituitary-dependent adrenal responses after endotoxin administration²³ provided early evidence that inflammatory stimuli can activate anti-inflammatory signals from the central nervous system (CNS). Subsequently, Besedovsky *et al.*²⁴ showed directly that inflammation in peripheral tissues alters neuronal signalling in the hypothalamus. Extensive work has identified a common molecular basis for communication, with cells from each system expressing signalling ligands and receptors from the other²⁵. For example, neurons in the CNS can synthesize and express TNF and IL-1, and these cytokines may participate in neuronal communication^{26,27}. This communication is bi-directional, because cytokines can activate hypothalamic-pituitary release of glucocorticoids and, in turn, glucocorticoids suppress further cytokine synthesis²⁸. In addition, cells of the immune system can produce neuropeptides (including endorphins), acetylcholine and other neurotransmitters.

The importance of the interaction between the nervous system and immune system signalling has been demonstrated recently in the development of pathological pain. Watkins and Maier²⁹ have proposed that cytokines produced by inflammatory and glial cells change neuronal excitability and that this link contributes directly to the development of intractable pain.

Cholinergic anti-inflammatory pathway

Our understanding of the basic mechanisms that regulate inflammation has been advanced by the identification of a neural mechanism that inhibits macrophage activation through parasympathetic outflow³⁰. Called the 'cholinergic anti-inflammatory pathway' because acetylcholine is the principle parasympathetic neurotransmitter, macrophages that are exposed to acetylcholine are effectively

Figure 1 The cholinergic anti-inflammatory pathway. Efferent activity in the vagus nerve leads to acetylcholine (ACh) release in organs of the reticuloendothelial system, including the liver, heart, spleen and gastrointestinal tract. Acetylcholine interacts with α -bungarotoxin-sensitive nicotinic receptors (ACh receptor) on tissue macrophages, which inhibit the release of TNF, IL-1, HMGB1 and other cytokines.



deactivated (Fig. 1). The vagus nerve (which was named for its wandering course) innervates the principal organs, including those that contain the reticuloendothelial system (liver, lung, spleen, kidneys and gut)³¹. Experimental activation of the cholinergic anti-inflammatory pathway by direct electrical stimulation of the efferent vagus nerve inhibits the synthesis of TNF in liver, spleen and heart, and attenuates serum concentrations of TNF during endotoxaemia^{30,32}. Vagotomy significantly exacerbates TNF responses to inflammatory stimuli and sensitizes animals to the lethal effects of endotoxin.

This 'hard-wired' connection between the nervous and immune systems functions as an anti-inflammatory mechanism in other models of systemic and local inflammation. Direct stimulation of the vagus nerve *in situ* inhibits proinflammatory cytokine synthesis in liver and cardiac tissue obtained from animals subjected to ischaemia-reperfusion by transient aortic clamping. Stimulation of either the right or the left cervical vagus nerves protects against the development of hypotension and inhibits serum TNF responses to ischaemia reperfusion³². The protection conferred by stimulation of the vagus nerve is dependent on the applied voltage and is associated with normalization of tachycardia during the reperfusion-induced hypotensive phase³². In a standardized model of experimental murine arthritis induced by the application of carrageenan, vagus nerve stimulation inhibits the inflammatory response and suppresses the development of paw swelling, indicating that the cholinergic anti-inflammatory pathway can inhibit localized inflammation specifically³³.

The molecular dovetail between the cholinergic nervous system and the innate immune system is a nicotinic, α -bungarotoxin-sensitive macrophage acetylcholine receptor³⁰. Exposure of human macrophages, but not peripheral blood monocytes, to nicotine or acetylcholine inhibits the release of TNF, IL-1 and IL-18 in response to endotoxin. Tissue macrophages, but not circulating monocytes, produce most of the TNF that appears systemically during an excessive inflammatory response. Interaction between the macrophage cholinergic receptor and its ligand inhibits the synthesis of proinflammatory cytokines (TNF, IL-1 and IL-18) but not anti-inflammatory cytokines (such as IL-10)³⁰. Acetylcholine inhibits the expression of TNF protein in macrophages, but not the induction of TNF messenger RNA levels, indicating that activation of

the cholinergic receptor transduces intracellular signals that inhibit cytokine synthesis at a post-transcriptional stage.

As compared with macrophages, monocytes are refractory to the cytokine-inhibiting effects of acetylcholine: only supraphysiological concentrations of cholinergic agonists inhibit cytokine synthesis in monocytes³⁰. The macrophage acetylcholine receptor is distinct from the muscarinic receptor activities identified on lymphocytes, peripheral blood mononuclear cells and alveolar macrophages^{34,35}. The exquisite sensitivity of macrophages to acetylcholine suggests that other non-neuronal cells that produce acetylcholine (such as epithelial cells, T lymphocytes and endothelial cells) might also participate in modulating the function of adjacent tissue macrophages^{36,37}.

Vagus nerve stimulation suppresses inflammation

Stimulation of efferent vagus nerve activity has been associated classically with slowing heart rate, induction of gastric motility, dilation of arterioles and constriction of pupils. Inhibition of the inflammatory response can now be added to this list (Fig. 1). From an oversimplified, teleological engineering perspective, there are many reasons why a neural-based anti-inflammatory pathway is advantageous. The diffusible anti-inflammatory network, which includes glucocorticoids, anti-inflammatory cytokines and other humoral mediators, is slow, distributed, non-integrated and dependent on concentration gradients. By contrast, the cholinergic anti-inflammatory pathway is discrete and localized in tissues where invasion and injury typically originate (Fig. 2).

As compared with the routine, biological pace of a typical, diffusible inflammatory response (hours to days), neural signalling is like lightning. This regulatory attribute is highly advantageous for containing immune activation at the crucial stages of a nascent response. Neural control of biological functions is short-lived: after a brief refractory period, responding cells can resume function as required in the absence of further neural input. Recovery of immune function after transient inhibition enables necessary local inflammatory responses to be mobilized during persisting threat or infection. The impact of sensitization or desensitization developing after, respectively, denervation or repeated neural firing to an inflammatory site has not been explored, but it would be predicted to influence anti-inflammatory function.

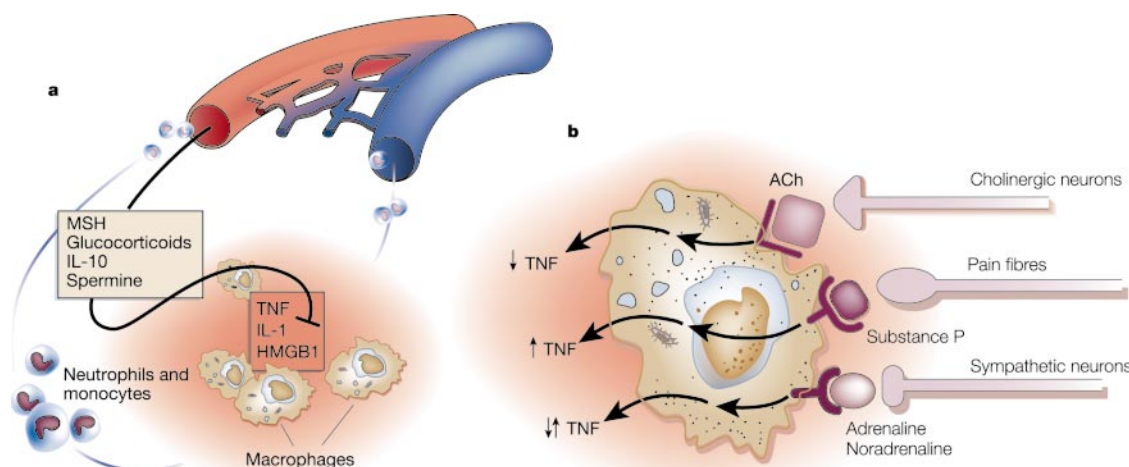


Figure 2 Diffusible versus neural anti-inflammatory pathways. **a**, Diffusible pathways. The circulation delivers inflammatory cells (monocytes and neutrophils) and cytokines to and from the inflammatory site; these responses are concentration gradient-dependent, slow and not integrated. Inflammatory products produced in the damaged tissue (TNF, IL-1, HMGB1) diffuse into the bloodstream, and anti-inflammatory hormones and cytokines (glucocorticoids, α -MSH, IL-10, spermine) diffuse into the zone. **b**, Neural pathways. Neural anti-inflammatory regulation of tissue macrophages is local, fast and integrated through the CNS. Acetylcholine inhibits the release of TNF from macrophages. Adrenaline and noradrenaline predominately inhibit TNF release but can, under some circumstances, stimulate TNF release. Substance P can stimulate cytokine synthesis to amplify the local inflammatory response and can also mediate pain.

Neural regulation of discrete, distributed, localized inflammatory sites provides a mechanism for integrating responses in real time. It is intriguing to consider that, in addition to the development of immunological memory, the involvement of the cholinergic anti-inflammatory pathway might also modulate processing events that promote neural memory of the peri-inflammatory events (that is, the 'hissing snake' or 'charging lion' that caused the wound and/or infection). Clark *et al.*³⁸ recently discovered that electrical stimulation of the vagus nerve in humans significantly enhanced word-recognition memory, indicating that memory formation and vagus nerve activity are closely linked.

Sensory function of vagus nerve signals in inflammation

The CNS receives sensory input from the immune system through both humoral and neural routes. Blalock^{39,40} originally suggested that the immune system functions as a 'sixth sense' that detects microbial invasion and produces molecules that relay this information to the brain. TNF and other immunological mediators can gain access to brain centres that are devoid of a blood–brain barrier in the circum-ventricular region. Indeed, the dorsal vagal complex, comprising the sensory nuclei of the solitary tract, the area postrema and the dorsal motor nucleus of the vagus, responds to increased circulating amounts of TNF by altering motor activity in the vagus nerve^{41–43}. This humoral route for communication between the immune system and the nervous system has been implicated in the development of fever, anorexia, activation of hypothalamic-pituitary responses to infection and injury, and other behavioural manifestations of illness.

Sensory innervation of immune organs by ascending fibres travelling in the vagus nerve, as well as by other pain and ascending sensory pathways, provides important input about the status of invasive and injurious challenges in distributed body compartments. Notably, these neural inflammation-sensing pathways can function at low thresholds of detection and can activate responses even when the inflammatory agents are present in tissues in quantities that are not high enough to reach the brain through the bloodstream. Watkins and colleagues^{44–47} have provided insight into the sensory role of afferent vagus nerve fibres by observing that vagotomy blunts the development of fever in animals exposed to intra-abdominal IL-1.

The afferent vagus pathway is activated by very low doses of endotoxin or IL-1; but higher doses of these agents can directly activate thermogenic responses through the humoral route to the brain⁴⁸. It is not completely clear how the vagus nerve 'detects' the presence of low doses of endotoxin or other inflammatory agents, but neurons in the vagus nerve express IL-1 receptor mRNA and discrete IL-1-binding sites have been identified on glomus cells in the vagus nerve proper^{41,49}.

Electrophysiological studies indicate that vagus nerve signals also can be activated by TNF, other cytokines, mechanoreceptors, chemoreceptors, temperature sensors and osmolarity sensors that might be activated at an inflammatory locus⁵⁰. Somatic sensory input into the CNS is organized somatotopically, such that sensory input from a discrete peripheral site is localized precisely in the ascending fibre pathways and brain. The first CNS synapse for afferent vagus signals lies in the nucleus tractus solitarius, and electrolytic lesioning of this region impairs the development of IL-1-induced fever⁵¹. Thus, inflammation-derived sensory input can be processed differentially in the brain, depending on the location of the inflammatory site and the nature of the sensory signal.

Reflex inhibition of inflammation

The inflammation-sensing and inflammation-suppressing functions outlined above provide the principal components of the inflammatory reflex (Fig. 3). The appearance of pathogenic organisms in a local wound, or at the site of epithelial barrier dysfunction, activates innate immune cells that release cytokines. These activate sensory fibres that ascend in the vagus nerve to synapse in the nucleus tractus solitarius. Increased efferent signals in the vagus nerve suppress peripheral cytokine release through macrophage nicotinic receptors and the cholinergic anti-inflammatory pathway. The 'inflammatory reflex' is described as localized, rapid and discrete; but it can also induce systemic humoral anti-inflammatory responses. This occurs because vagus nerve activity can be relayed to the medullary reticular formation, to the locus ceruleus and to the hypothalamus, leading to increased release of ACTH from the anterior pituitary.

Increased cytokine production in tissues causes pain, providing another mechanism for transferring information from the immune

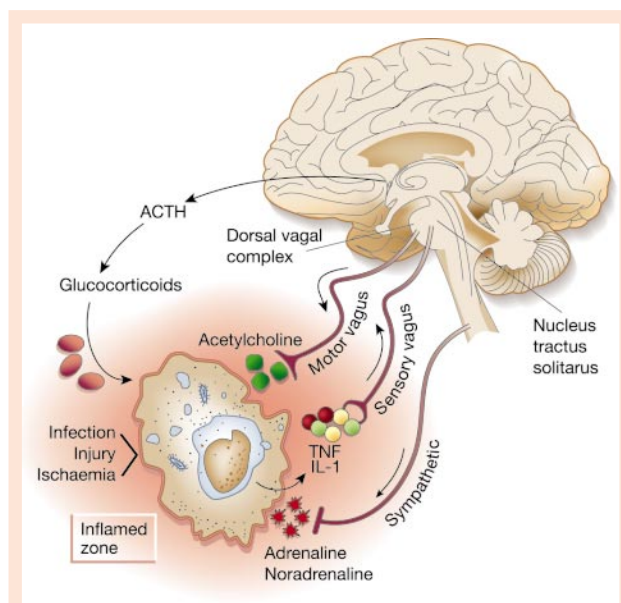


Figure 3 Wiring of the inflammatory reflex. Inflammatory products produced in damaged tissues activate afferent signals that are relayed to the nucleus tractus solitarius; subsequent activation of vagus efferent activity inhibits cytokine synthesis through the cholinergic anti-inflammatory pathway ('the inflammatory reflex'). Information can also be relayed to the hypothalamus and the dorsal vagal complex to stimulate the release of ACTH, thereby activating the humoral anti-inflammatory pathway. Activation of the sympathetic outflow by flight-or-fight responses or pain, or through direct signalling, can increase local concentrations of adrenaline and noradrenaline, which can suppress inflammation further.

system to the brain. This information can be relayed to other brain centres that influence motor output in the vagus nerve. Pain and stress can activate the flight-or-fight responses, and the resultant increase of adrenaline and noradrenaline also can inhibit macrophage activation and suppress synthesis of TNF and other cytokines^{13,52,53}. High sympathetic activity and resultant increases in catecholamines stimulate the β -adrenergic-receptor-dependent release of IL-10, a potent anti-inflammatory cytokine, from monocytes^{11,54}. Thus, the anti-inflammatory effects of the sympathetic and parasympathetic nervous systems seem to be synergistic in this setting.

Classical teaching stresses that actions of the sympathetic and parasympathetic nervous systems are usually in opposition. But in many situations the two systems function synergistically. For example, simultaneous stimulation of both sympathetic and vagus nerves produces a higher increase in cardiac output than does isolated stimulation of either nerve alone⁵⁵. Flight-or-fight activation of sympathetic responses also stimulates increased vagus nerve output. The combined action of these neural systems is significantly anti-inflammatory and is positioned anatomically to constrain local inflammation by preventing spillover of potentially lethal toxins into the circulation through both local (neural) and systemic (humoral) anti-inflammatory mechanisms.

Implications of the inflammatory reflex

Knowledge of the inflammatory reflex and the cholinergic anti-inflammatory pathway is yielding insight into both physiological pathways and therapeutic strategies (Fig. 4). For example, it may be possible to activate neural anti-inflammatory mechanisms using small molecules that initiate signals in proximal components of the pathway in the CNS. One such molecule is CNI-1493, a tetravalent guanlylhydrazone that was originally described as an inhibitor of macrophage activation and TNF release^{56,57}.

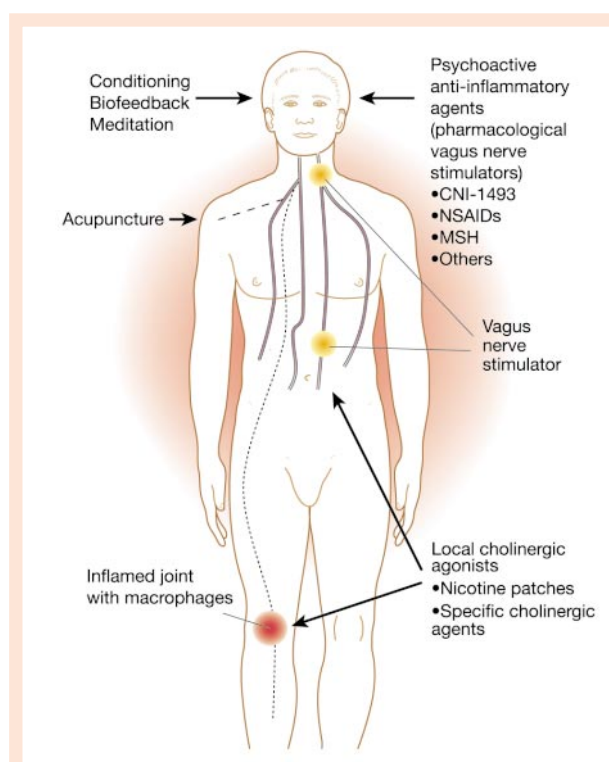


Figure 4 Targeting therapies to the cholinergic anti-inflammatory pathway. The physiological basis of the cholinergic anti-inflammatory pathway could guide the development of therapies based on either modulating the activity of the vagus nerve or targeting specific components of the pathway. For example, biofeedback, conditioning, meditation, hypnosis or acupuncture could be potentially used to modulate vagus output, 'psychoactive' drugs could be tailor-made to increase vagus output ('pharmacological vagus nerve stimulators'; NSAIDs, non-steroidal anti-inflammatory drugs), and other agents could be used to target macrophage cholinergic receptors in the periphery. Unbroken lines represent known vagus nerve pathways; dotted lines are hypothetical.

CNI-1493 inhibits TNF synthesis and inflammatory responses in animal models of local and systemic inflammation⁵⁸. It also significantly reduced disease severity in a small clinical trial of severe Crohn's disease and is currently being evaluated in a large phase II trial of Crohn's disease⁵⁹. Unexpectedly, recent evidence has shown that the TNF-suppressing activities of CNI-1493 *in vivo* are dependent on the cholinergic anti-inflammatory pathway, and that CNI-1493 functions as a pharmacological stimulator of the vagus nerve^{32,60}. Intracerebral application of small doses of CNI-1493 significantly inhibited peripheral TNF synthesis, and intact vagus nerves were required to prevent increases in serum TNF. The mechanism through which CNI-1493 activates the vagus nerve is unknown, but increased vagus nerve firing has been observed after either intracerebral or intravenous administration of CNI-1493 — an effect that seems to be dependent on specific CNS receptors³³.

It is likely that other experimental and clinically approved therapeutic agents suppress peripheral inflammation by activating pathways in the CNS. Small doses of α -MSH applied intracerebrally inhibited pulmonary myeloperoxidase activity in mice exposed to endotoxin⁶¹ and suppressed the development of intradermal oedema induced by exposure to TNF or IL-1 (ref. 62). Specific anti-inflammatory responses have been observed in response to intracerebral application of salicylates, but not dexamethasone⁶³. The cardiac anti-arrhythmic drug amiodarone has been identified as an inhibitor of TNF synthesis in monocytes *in vitro*⁶⁴, but it also functions as a potent stimulator of vagus nerve activity⁶⁵. Systemic administration of the non-steroidal anti-inflammatory drugs aspirin, indomethacin and

ibuprofen substantially increases vagus nerve activity⁶⁶. Although this vagus nerve response had been studied in the context of increasing gastric acidity and ulcer formation, knowledge of the cholinergic anti-inflammatory pathways raises the possibility that the vagus-nerve-stimulating activity of these agents may also contribute to their anti-inflammatory action. A better understanding of the CNS receptors, pathways and neural mechanisms that activate the vagus nerve to inhibit production of TNF should facilitate development of this pharmacological 'vagus-nerve-stimulating' approach.

Another experimental therapeutic approach is based on direct electrical stimulation of the vagus nerve. So far, more than 10,000 individuals have received implantable vagus nerve stimulators for the treatment of epilepsy^{67,68}. Vagus nerve stimulation in humans with small, pacemaker-like devices is safe, well tolerated and not associated with increased rates of infection. But the immunological effects of this approach have not been reported and, indeed, it will be interesting to assess whether stimulating the vagus nerve in humans modulates TNF synthesis and inflammation. In place of implantable devices, it should be possible to develop pharmacological approaches that target the peripheral macrophage receptor to inhibit TNF synthesis. A precedent for this approach has been already achieved in the clinic, because nicotine administration is significantly efficacious in reducing the severity of ulcerative colitis⁶⁹. Other preclinical studies using standard murine models of diabetes have shown that nicotine reduces the incidence of diabetes by reducing pancreatic concentrations of TNF and other cytokines⁷⁰. Unanticipated activities of the cholinergic anti-inflammatory pathway in inflammatory disease and in non-immune cells might be determined by further studies.

Some of the earliest studies of the nervous system and inflammation examined the effects of pavlovian conditioning on intra-abdominal inflammatory responses⁷¹. Behavioural conditioning using models of learned association can reproducibly influence acute inflammatory responses and alter the course of experimental inflammatory diseases in animals and humans^{72–74}. Hypnosis and meditation can significantly increase vagus nerve output and have been observed to inhibit immediate-type and delayed-type hypersensitivity responses^{75,76}. Biofeedback and acupuncture have been used to modulate vagus nerve activity to alter bowel function, gastric acidity and heart rate^{77,78}. Each of these approaches has been used to reduce experimental inflammation, but the relationships between vagus nerve activity and anti-inflammatory action had not been defined previously. Autonomic dysfunction occurs as a classical complication of rheumatoid arthritis, diabetes and other autoimmune disorders^{79–81}. It is now intriguing to consider whether vagus nerve dysfunction underlies the progression of inflammation, owing to impairment of the cholinergic anti-inflammatory pathway. It is reasonable to propose that, one day, the rational modulation of vagus nerve activity using these or other approaches may provide a therapeutic advantage for inflammatory disease. □

doi:10.1038/nature01321

- Tracey, K. J. *et al.* Anti-cachectin/TNF monoclonal antibodies prevent septic shock during lethal bacteraemia. *Nature* **330**, 662–664 (1987).
- Tracey, K. J. *et al.* Shock and tissue injury induced by recombinant human cachectin. *Science* **234**, 470–474 (1986).
- Wang, H. *et al.* HMGB-1 as a late mediator of endotoxin lethality in mice. *Science* **285**, 248–251 (1999).
- Tracey, K. J., Vlassara, H. & Cerami, A. Cachectin/tumour necrosis factor. *Lancet* **i**, 1122–1126 (1989).
- Tracey, K. J. & Abraham, E. From mouse to man: or what have we learned about cytokine-based anti-inflammatory therapies? *Shock* **11**, 224–225 (1999).
- Andersson, U. *et al.* High mobility group 1 protein (HMGB-1) stimulates proinflammatory cytokine synthesis in human monocytes. *J. Exp. Med.* **192**, 565–570 (2000).
- Ulloa, L. *et al.* Ethyl pyruvate prevents lethality in mice with established lethal sepsis and systemic inflammation. *Proc. Natl Acad. Sci. USA* **99**, 12351–12356 (2002).
- Wang, H., Yang, H., Czura, C. J., Sama, A. E. & Tracey, K. J. HMGB1 as a late mediator of lethal systemic inflammation. *Am. J. Respir. Crit. Care Med.* **164**, 1768–1773 (2001).
- Lantz, M., Gullberg, U., Nilsson, E. & Olsson, I. Characterization *in vitro* of a human tumor necrosis factor-binding protein. A soluble form of a tumor necrosis factor receptor. *J. Clin. Invest.* **86**, 1396 (1990).
- Tsunawaki, S., Sporn, M., Ding, A. & Nathan, C. Deactivation of macrophages by transforming growth factor- β . *Nature* **334**, 260–262 (1988).
- Van der, P. T., Coyle, S. M., Barbosa, K., Braxton, C. C. & Lowry, S. F. Epinephrine inhibits tumor

- necrosis factor- α and potentiates interleukin 10 production during human endotoxemia. *J. Clin. Invest.* **97**, 713–719 (1996).
- Scheinman, R. I., Cogswell, P. C., Lofquist, A. K. & Baldwin, A. S. Jr Role of transcriptional activation of I κ B α in mediation of immunosuppression by glucocorticoids. *Science* **270**, 283–286 (1995).
- Chrousos, G. P. The stress response and immune function: clinical implications. The 1999 Novera H. Spector Lecture. *Ann. NY Acad. Sci.* **917**, 38–67 (2000).
- Madden, K. S., Sanders, V. M. & Felten, D. L. Catecholamine influences and sympathetic neural modulation of immune responsiveness. *Annu. Rev. Pharmacol. Toxicol.* **35**, 417–448 (1995).
- Zhang, M., Borovikova, L. V., Wang, H., Metz, C. & Tracey, K. J. Spermine inhibition of monocyte activation and inflammation. *Mol. Med.* **5**, 595–605 (1999).
- Bertini, R., Bianchi, M. & Ghezzi, P. Adrenalectomy sensitizes mice to the lethal effects of interleukin 1 and tumor necrosis factor. *J. Exp. Med.* **167**, 1708–1712 (1988).
- Butler, L. D. *et al.* Neuroendocrine regulation of *in vivo* cytokine production and effects: I. *In vivo* regulatory networks involving the neuroendocrine system, interleukin-1 and tumor necrosis factor- α . *J. Neuroimmunol.* **24**, 143–153 (1989).
- Bloom, O. *et al.* Hypophysectomy, high tumor necrosis factor levels, and hemoglobinemia in lethal endotoxemic shock. *Shock* **10**, 395–400 (1998).
- Sternberg, E. M. *et al.* Inflammatory mediator-induced hypothalamic-pituitary-adrenal axis activation is defective in streptococcal cell wall arthritis-susceptible Lewis rats. *Proc. Natl Acad. Sci. USA* **86**, 2374–2378 (1989).
- Webster, J. I., Tonelli, L. & Sternberg, E. M. Neuroendocrine regulation of immunity. *Annu. Rev. Immunol.* **20**, 125–163 (2002).
- Davidson, N. J. *et al.* T helper cell 1-type CD4⁺ T cells, but not B cells, mediate colitis in interleukin 10-deficient mice. *J. Exp. Med.* **184**, 241–251 (1996).
- Johansson, A. C., Hansson, A. S., Nandakumar, K. S., Backlund, J. & Holmdahl, R. IL-10-deficient B10.Q mice develop more severe collagen-induced arthritis, but are protected from arthritis induced with anti-type II collagen antibodies. *J. Immunol.* **167**, 3505–3512 (2001).
- Wexler BC, Dolgin AE & Tryczynski EW. Effects of a bacterial polysaccharide (Pirumen) on the pituitary-adrenal axis: adrenal ascorbic acid, cholesterol and histologic alterations. *Endocrinology* **61**, 300–308 (1957).
- Besedovsky, H., Sorkin, E., Felix, D. & Haas, H. Hypothalamic changes during the immune response. *Eur. J. Immunol.* **7**, 323–325 (1977).
- Blalock, J. E. A molecular basis for bidirectional communication between the immune and neuroendocrine systems. *Physiol. Rev.* **69**, 1–32 (1989).
- Breder, C. D., Dinarello, C. A. & Saper, C. B. Interleukin-1 immunoreactive innervation of the human hypothalamus. *Science* **240**, 321–324 (1988).
- Breder, C. D. *et al.* Regional induction of tumor necrosis factor α expression in the mouse brain after systemic lipopolysaccharide administration. *Proc. Natl Acad. Sci. USA* **91**, 11393–11397 (1994).
- Besedovsky, H., del Rey, A., Sorkin, E. & Dinarello, C. A. Immunoregulatory feedback between interleukin-1 and glucocorticoid hormones. *Science* **233**, 652–654 (1986).
- Watkins, L. R. & Maier, S. F. Beyond neurons: evidence that immune and glial cells contribute to pathological pain states. *Physiol. Rev.* **82**, 981–1011 (2002).
- Borovikova, L. V. *et al.* Vagus nerve stimulation attenuates the systemic inflammatory response to endotoxin. *Nature* **405**, 458–462 (2000).
- Bellinger, D. L., Lorton, D., Lubahn, C. & Felten, D. L. in *Psychoneuroimmunology* (eds Ader R., Felten, D. L. & Cohen, N.) 55–112 (Academic, San Diego, 2001).
- Bernik, T. R. *et al.* Pharmacological stimulation of the cholinergic antiinflammatory pathway. *J. Exp. Med.* **195**, 781–788 (2002).
- Borovikova, L. V. *et al.* Role of vagus nerve signaling in CNI-1493-mediated suppression of acute inflammation. *Auton. Neurosci.* **85**, 141–147 (2000).
- Sato, K. Z. *et al.* Diversity of mRNA expression for muscarinic acetylcholine receptor subtypes and neuronal nicotinic acetylcholine receptor subunits in human mononuclear leukocytes and leukemic cell lines. *Neurosci. Lett.* **266**, 17–20 (1999).
- Sato, E., Koyama, S., Okubo, Y., Kubo, K. & Sekiguchi, M. Acetylcholine stimulates alveolar macrophages to release inflammatory cell chemotactic activity. *Am. J. Physiol.* **274**, L970–L979 (1998).
- Wessler, I., Kirkpatrick, C. J. & Racke, K. Non-neuronal acetylcholine, a locally acting molecule, widely distributed in biological systems: expression and function in humans. *Pharmacol. Ther.* **77**, 59–79 (1998).
- Kawashima, K. & Fujii, T. Extraneuronal cholinergic system in lymphocytes. *Pharmacol. Ther.* **86**, 29–48 (2000).
- Clark, K. B., Naritoku, D. K., Smith, D. C., Browning, R. A. & Jensen, R. A. Enhanced recognition memory following vagus nerve stimulation in human subjects. *Nature Neurosci.* **2**, 94–98 (1999).
- Blalock, J. E. The immune system as a sensory organ. *J. Immunol.* **132**, 1067–1070 (1984).
- Blalock, J. E. Shared ligands and receptors as a molecular mechanism for communication between the immune and neuroendocrine systems. *Ann. NY Acad. Sci.* **741**, 292–298 (1994).
- Goehler, L. E. *et al.* Vagal immune-to-brain communication: a visceral chemosensory pathway. *Auton. Neurosci.* **85**, 49–59 (2000).
- Hermann, G. E., Emch, G. S., Tovar, C. A. & Rogers, R. C. c-Fos generation in the dorsal vagal complex after systemic endotoxin is not dependent on the vagus nerve. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **280**, R289–R299 (2001).
- Emch, G. S., Hermann, G. E. & Rogers, R. C. TNF- α activates solitary nucleus neurons responsive to gastric distension. *Am. J. Physiol. Gastrointest. Liver Physiol.* **279**, G582–G586 (2000).
- Watkins, L. R. & Maier, S. F. Implications of immune-to-brain communication for sickness and pain. *Proc. Natl Acad. Sci. USA* **96**, 7710–7713 (1999).
- Watkins, L. R. *et al.* Blockade of interleukin-1 induced hyperthermia by subdiaphragmatic vagotomy: evidence for vagal mediation of immune-brain communication. *Neurosci. Lett.* **183**, 27–31 (1995).
- Hansen, M. K. *et al.* Effects of vagotomy on lipopolysaccharide-induced brain interleukin-1 β protein in rats. *Auton. Neurosci.* **85**, 119–126 (2000).
- Hansen, M. K., O'Connor, K. A., Goehler, L. E., Watkins, L. R. & Maier, S. F. The contribution of the vagus nerve in interleukin-1 β -induced fever is dependent on dose. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **280**, R929–R934 (2001).
- Romanovsky, A. A. Thermoregulatory manifestations of systemic inflammation: lessons from vagotomy. *Auton. Neurosci. Basic Clin.* **85**, 39–48 (2000).
- Goehler, L. E. *et al.* Vagal paranglia bind biotinylated interleukin-1 receptor antagonist: a possible mechanism for immune-to-brain communication. *Brain Res. Bull.* **43**, 357–364 (1997).

50. Berthoud, H. R. & Neuhuber, W. L. Functional and chemical anatomy of the afferent vagal system. *Auton. Neurosci.* **85**, 1–17 (2000).
51. Gordon, F. J. Effect of nucleus tractus solitarius lesions on fever produced by interleukin-1 β . *Auton. Neurosci.* **85**, 102–110 (2000).
52. Molina, P. E., Bagby, G. J. & Stahls, P. Hemorrhage alters neuroendocrine, hemodynamic, and compartment-specific TNF responses to LPS. *Shock* **16**, 459–465 (2001).
53. Molina, P. E. Noradrenergic inhibition of TNF upregulation in hemorrhagic shock. *Neuroimmunomodulation* **9**, 125–133 (2001).
54. Woiciechowsky, C. *et al.* Sympathetic activation triggers systemic interleukin-10 release in immunodepression induced by brain injury. *Nature Med.* **4**, 808–813 (1998).
55. Koizumi, K., Terui, N., Kollai, M. & Brooks, C. M. Functional significance of coactivation of vagal and sympathetic cardiac nerves. *Proc. Natl Acad. Sci. USA* **79**, 2116–2120 (1982).
56. Bianchi, M. *et al.* Suppression of proinflammatory cytokines in monocytes by a tetravalent guanylylhydrazide. *J. Exp. Med.* **183**, 927–936 (1996).
57. Bianchi, M. *et al.* An inhibitor of macrophage arginine transport and nitric oxide production (CNI-1493) prevents acute inflammation and endotoxin lethality. *Mol. Med.* **1**, 254–266 (1995).
58. Tracey, K. J. Suppression of TNF and other proinflammatory cytokines by the tetravalent guanylylhydrazide CNI-1493. *Prog. Clin. Biol. Res.* **397**, 335–343 (1998).
59. Hommes, D. *et al.* Inhibition of stress-activated MAP kinases induces clinical improvement in moderate to severe Crohn's disease. *Gastroenterology* **122**, 7–14 (2002).
60. Tracey, K. J., Czura, C. J. & Ivanova, S. Mind over immunity. *FASEB J.* **15**, 1575–1576 (2001).
61. Delgado, H. R. *et al.* Inhibition of systemic inflammation by central action of the neuropeptide α -melanocyte-stimulating hormone. *Neuroimmunomodulation* **6**, 187–192 (1999).
62. Ceriani, G., Macaluso, A., Catania, A. & Lipton, J. M. Central neurogenic antiinflammatory action of α -MSH: modulation of peripheral inflammation induced by cytokines and other mediators of inflammation. *Neuroendocrinology* **59**, 138–143 (1994).
63. Catania, A., Arnold, J., Macaluso, A., Hiltz, M. E. & Lipton, J. M. Inhibition of acute inflammation in the periphery by central action of salicylates. *Proc. Natl Acad. Sci. USA* **88**, 8544–8547 (1991).
64. Matsumori, A., Ono, K., Nishio, R., Nose, Y. & Sasayama, S. Amiodarone inhibits production of tumor necrosis factor- α by human mononuclear cells: a possible mechanism for its effect in heart failure. *Circulation* **96**, 1386–1389 (1997).
65. Dias, D. S. *et al.* Opposite effects of iv amiodarone on cardiovascular vagal and sympathetic efferent activities in rats. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **283**, R543–R548 (2002).
66. Arai, I., Hirose, H., Muramatsu, M., Okuyama, S. & Aihara, H. Possible involvement of non-steroidal anti-inflammatory drugs in vagal-mediated gastric acid secretion in rats. *Jpn. J. Pharmacol.* **37**, 91–99 (1985).
67. Ben Menachem, E. Vagus nerve stimulation, side effects, and long-term safety. *J. Clin. Neurophysiol.* **18**, 415–418 (2001).
68. Schachter, S. C. Vagus nerve stimulation: where are we? *Curr. Opin. Neurol.* **15**, 201–206 (2002).
69. Pullan, R. D. *et al.* Transdermal nicotine for active ulcerative colitis. *N. Engl. J. Med.* **330**, 811–815 (1994).
70. Mabley, J. G., Pacher, P., Southan, G. J., Salzman, A. L. & Szabo, C. Nicotine reduces the incidence of type I diabetes in mice. *J. Pharmacol. Exp. Ther.* **300**, 876–881 (2002).
71. Metal'nikov, S. a. V. C. Role des reflexes conditionnels dans l'immunité. *Ann. Inst. Pasteur* **40**, 893–900 (1926).
72. Madden, K. S. & Felten, D. L. Experimental basis for neural-immune interactions. *Physiol. Rev.* **75**, 77–106 (1995).
73. Ader R. & Cohen, N. in *Psychoneuroimmunology* (eds Ader R., Felten, D. L. & Cohen, N.) 3–34 (Academic, San Diego, 2001).
74. Exton, M. S. *et al.* Pavlovian conditioning of immune function: animal investigation and the challenge of human application. *Behav. Brain Res.* **110**, 129–141 (2000).
75. Black, S. Inhibition of immediate-type hypersensitivity response by direct suggestion under hypnosis. *Br. Med. J.* **1**, 925–929 (1963).
76. Zachariae, R. in *Psychoneuroimmunology* (eds Ader R., Felten, D. L. & Cohen, N.) 133–160 (Academic, San Diego, 2001).
77. Noguchi, E. & Hayashi, H. Increases in gastric acidity in response to electroacupuncture stimulation of the hindlimb of anesthetized rats. *Jpn. J. Physiol.* **46**, 53–58 (1996).
78. Lux, G. *et al.* Acupuncture inhibits vagal gastric acid secretion stimulated by sham feeding in healthy subjects. *Gut* **35**, 1026–1029 (1994).
79. Toussiot, E., Serratrice, G. & Valentin, P. Autonomic nervous system involvement in rheumatoid arthritis. 50 cases. *J. Rheumatol.* **20**, 1508–1514 (1993).
80. Tan, J., Akin, S., Beyazova, M., Sepici, V. & Tan, E. Sympathetic skin response and R-R interval variation in rheumatoid arthritis. Two simple tests for the assessment of autonomic function. *Am. J. Phys. Med. Rehabil.* **72**, 196–203 (1993).
81. Edmonds, M. E., Jones, T. C., Saunders, W. A. & Sturrock, R. D. Autonomic neuropathy in rheumatoid arthritis. *Br. Med. J.* **2**, 173–175 (1979).

Acknowledgements

Supported in part by grants from the National Institutes of Health (National Institute of General Medical Sciences) and the Defense Advanced Research Projects Agency (DARPA). The author is grateful for the thoughtful suggestions from C. Czura, M. Fink, S. Friedman, C. Nathan and B. Sherry.

Inflammation and cancer

Lisa M. Coussens*†§ & Zena Werb‡§

*Cancer Research Institute, Departments of †Pathology and ‡Anatomy, and the §UCSF Comprehensive Cancer Center, University of California, San Francisco, California 94143 USA (e-mail: coussens@cc.ucsf.edu; zena@itsa.ucsf.edu)

Recent data have expanded the concept that inflammation is a critical component of tumour progression. Many cancers arise from sites of infection, chronic irritation and inflammation. It is now becoming clear that the tumour microenvironment, which is largely orchestrated by inflammatory cells, is an indispensable participant in the neoplastic process, fostering proliferation, survival and migration. In addition, tumour cells have co-opted some of the signalling molecules of the innate immune system, such as selectins, chemokines and their receptors for invasion, migration and metastasis. These insights are fostering new anti-inflammatory therapeutic approaches to cancer development.

The functional relationship between inflammation and cancer is not new. In 1863, Virchow hypothesized that the origin of cancer was at sites of chronic inflammation, in part based on his hypothesis that some classes of irritants, together with the tissue injury and ensuing inflammation they cause, enhance cell proliferation¹. Although it is now clear that proliferation of cells alone does not cause cancer, sustained cell proliferation in an environment rich in inflammatory cells, growth factors, activated stroma, and DNA-damage-promoting agents, certainly potentiates and/or promotes neoplastic risk. During tissue injury associated with wounding, cell proliferation is enhanced while the tissue regenerates; proliferation and inflammation subside after the assaulting agent is removed or the repair completed. In contrast, proliferating cells that sustain DNA damage and/or mutagenic assault (for example, initiated cells) continue to proliferate in microenvironments rich in inflammatory cells and growth/survival factors that support their growth. In a sense, tumours act as wounds that fail to heal².

Today, the causal relationship between inflammation, innate immunity and cancer is more widely accepted; however, many of the molecular and cellular mechanisms mediating this relationship remain unresolved — these are the focus of this review. Furthermore, tumour cells may usurp key mechanisms by which inflammation interfaces with cancers, to further their colonization of the host. Although the acquired immune response to cancer is intimately related to the inflammatory response, this topic is beyond the scope of this article, but readers are referred to several excellent reviews^{3,4}.

An overview of inflammation

To understand the role of inflammation in the evolution of cancer, it is important to understand what inflammation is and how it contributes to physiological and pathological processes such as wound healing and infection (Fig. 1). In response to tissue injury, a multifactorial network of chemical signals initiate and maintain a host response designed to 'heal' the afflicted tissue. This involves activation and directed migration of leukocytes (neutrophils, monocytes and eosinophils) from the venous system to sites of damage (Box 1), and tissue mast cells also have a significant role. For neutrophils, a four-step mechanism is believed to coordinate recruitment of these inflammatory cells to sites of tissue injury and to the provisional extracellular matrix (ECM) that forms a scaffolding upon which fibroblast and endothelial cells proliferate and migrate, thus providing a nidus for

reconstitution of the normal microenvironment⁵. These steps involve: activation of members of the selectin family of adhesion molecules (L-, P-, and E-selectin) that facilitate rolling along the vascular endothelium; triggering of signals that activate and upregulate leukocyte integrins mediated by cytokines and leukocyte-activating molecules; immobilization of neutrophils on the surface of the vascular endothelium by means of tight adhesion through $\alpha_4\beta_1$ and $\alpha_4\beta_7$ integrins binding to endothelial vascular cell-adhesion molecule-1 (VCAM-1) and MadCAM-1, respectively; and transmigration through the endothelium to sites of injury, presumably facilitated by extracellular proteases, such as matrix metalloproteinases (MMPs).

A family of chemotactic cytokines, named chemokines, which possess a relatively high degree of specificity for chemoattraction of specific leukocyte populations^{1,6,7}, recruits downstream effector cells and dictates the natural evolution of the inflammatory response. The profile of cytokine/chemokines persisting at an inflammatory site is important in the development of chronic disease. The pro-inflammatory cytokine TNF- α (tumour necrosis factor- α) controls inflammatory cell populations as well as mediating many of the other aspects of the inflammatory process. TGF- β 1 is also important, both positively and negatively influencing the processes of inflammation and repair⁸. The key concept is that normal inflammation — for example, inflammation associated with wound healing — is usually self-limiting; however, dysregulation of any of the converging factors can lead to abnormalities and ultimately, pathogenesis — this seems to be the case during neoplastic progression.

Neutrophils (and sometimes eosinophils) are the first recruited effectors of the acute inflammatory response. Monocytes, which differentiate into macrophages in tissues, are next to migrate to the site of tissue injury, guided by chemotactic factors. Once activated, macrophages are the main source of growth factors and cytokines, which profoundly affect endothelial, epithelial and mesenchymal cells in the local microenvironment. Mast cells are also important in acute inflammation owing to their release of stored and newly synthesized inflammatory mediators, such as histamine, cytokines and proteases complexed to highly sulphated proteoglycans, as well as lipid mediators.

Inflammation and neoplastic progression

Peyton Rous was the first to recognize that cancers develop from "subthreshold neoplastic states" caused by viral or chemical carcinogens that induce somatic changes^{9,10}. These states, now known as 'initiation', involve DNA alter-

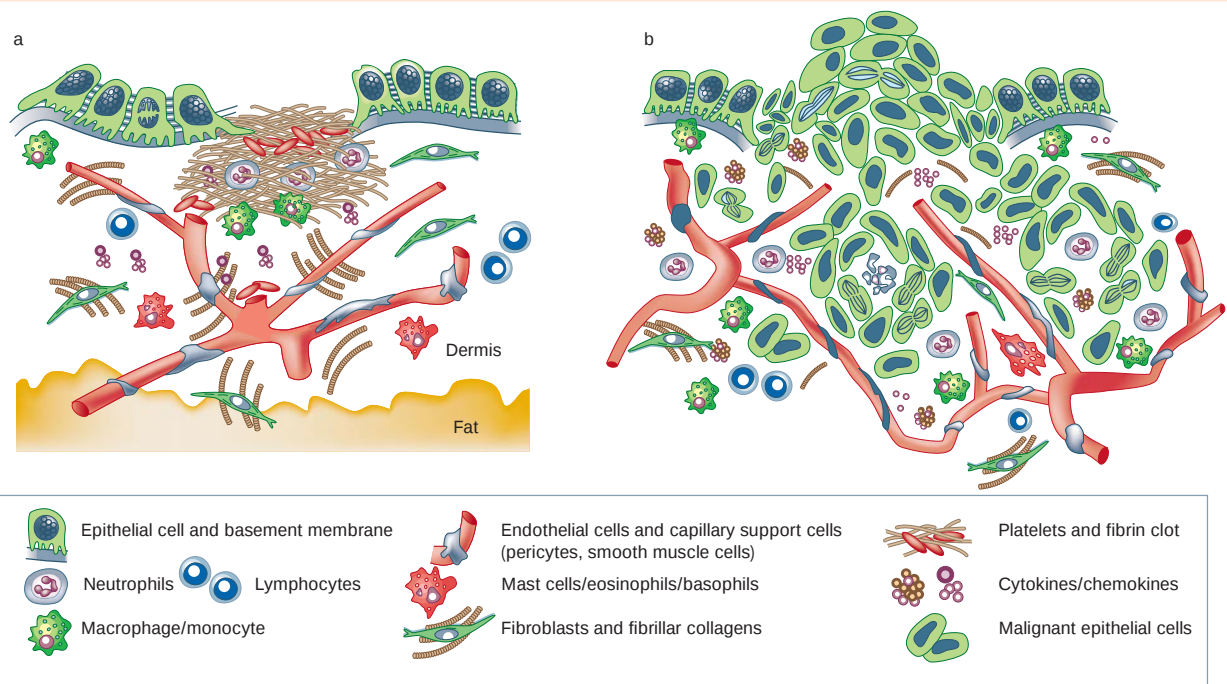


Figure 1 Wound healing versus invasive tumour growth. **a**, Normal tissues have a highly organized and segregated architecture. Epithelial cells sit atop a basement membrane separated from the vascularized stroma (dermis) compartment. Upon wounding or tissue assault, platelets are activated and form a haemostatic plug where they release vasoactive mediators that regulate vascular permeability, influx of serum fibrinogen, and formation of the fibrin clot. Chemotactic factors such as transforming growth factor- β and platelet-derived growth factor, derived from activated platelets, initiate granulation tissue formation, activation of fibroblasts, and induction and activation of proteolytic enzymes necessary for remodelling of the extracellular matrix (for example, matrix metalloproteinases and urokinase-type plasminogen activator). In combination, granulocytes, monocytes and fibroblasts are recruited, the venous network restored, and re-epithelialization across the wound occurs. Epithelial and stromal cell types engage in a reciprocal signalling dialogue to facilitate healing. Once the wound is healed, the reciprocal signalling subsides. **b**, Invasive carcinomas are less organized. Neoplasia-associated angiogenesis and lymphangiogenesis produces a chaotic vascular organization of blood vessels and lymphatics where neoplastic cells interact with other cell types (mesenchymal, haematopoietic and lymphoid) and a remodelled extracellular matrix. Although the vascular network is not disrupted in the same way during neoplastic progression as it is during wounding, many reciprocal interactions occur in parallel. Neoplastic cells produce an array of cytokines and chemokines that are mitogenic and/or chemoattractants for granulocytes, mast cells, monocytes/macrophages, fibroblasts and endothelial cells. In addition, activated fibroblasts and infiltrating inflammatory cells secrete proteolytic enzymes, cytokines and chemokines, which are mitogenic for neoplastic cells, as well as endothelial cells involved in neoangiogenesis and lymphangiogenesis. These factors potentiate tumour growth, stimulate angiogenesis, induce fibroblast migration and maturation, and enable metastatic spread via engagement with either the venous or lymphatic networks.

ations, are irreversible and can persist in otherwise normal tissue indefinitely until the occurrence of a second type of stimulation (now referred to as 'promotion'). Promotion can result from exposure of initiated cells to chemical irritants, such as phorbol esters, factors released at the site of wounding, partial organ resection, hormones or chronic irritation and inflammation (Fig. 1). Functionally, many promoters, whether directly or indirectly, induce cell proliferation, recruit inflammatory cells, increase production of reactive oxygen species leading to oxidative DNA damage, and reduce DNA repair. Subversion of cell death and/or repair programmes occurs in chronically inflamed tissues, thus resulting in DNA replication and proliferation of cells that have lost normal growth control. Normal inflammation is self-limiting, because the production of anti-inflammatory cytokines follows the pro-inflammatory cytokines closely (Fig. 2). However, chronic inflammation seems to be due to persistence of the initiating factors or a failure of mechanisms required for resolving the inflammatory response. Why does the inflammatory response to tumours persist?

Inflammatory cell component of tumours

Tumour cells produce various cytokines and chemokines that attract leukocytes. The inflammatory component of a developing neoplasm may include a diverse leukocyte population — for example, neutrophils, dendritic cells, macrophages, eosinophils and mast cells, as

well as lymphocytes — all of which are capable of producing an assorted array of cytokines, cytotoxic mediators including reactive oxygen species, serine and cysteine proteases, MMPs and membrane-perforating agents, and soluble mediators of cell killing, such as TNF- α , interleukins and interferons (IFNs)^{11,12}.

Monocytes, in the presence of granulocyte-macrophage colony-stimulating factor (GM-CSF) and interleukin (IL)-4, differentiate into immature dendritic cells¹³. Dendritic cells migrate into inflamed peripheral tissue where they capture antigens and, after maturation, migrate to lymph nodes to stimulate T-lymphocyte activation. Soluble factors such as IL-6 and CSF-1, derived from neoplastic cells, push myeloid precursors towards a macrophage-like phenotype¹⁴. Interestingly, dendritic cells found in neoplastic infiltrates are frequently immature and defective in T-cell stimulatory capacity.

Tumour-associated macrophages (TAMs) are a significant component of inflammatory infiltrates in neoplastic tissues and are derived from monocytes that are recruited largely by monocyte chemotactic protein (MCP) chemokines. TAMs have a dual role in neoplasms — although they may kill neoplastic cells following activation by IL-2, interferon and IL-12 (refs 15, 16), TAMs produce a number of potent angiogenic and lymphangiogenic growth factors, cytokines and proteases, all of which are mediators that potentiate neoplastic progression¹⁷. TAMs and tumour cells also produce IL-10, which effectively blunts the anti-tumour response by cytotoxic

Box 1

Wound healing as an example of physiological inflammation

Cellular components

Platelet activation and aggregation, in addition to accelerating coagulation, provide a bolus of secreted proteins and α -granule contents to the immediate area, all of which help initiate and accelerate the inflammatory response by the host. Examples of such secreted proteins include arachidonic acid metabolites, heparin, serotonin, thrombin, coagulation factors (factor V), adhesive proteins (fibrinogen and von Willebrand factor), plasma proteins (immunoglobulin- γ and albumin), cell growth factors (platelet-derived growth factor (PDGF), platelet-derived angiogenesis factor, transforming growth factor- α (TGF- α), TGF- β and basic fibroblast growth factor (bFGF)), enzymes (heparanase and factor XIII) and protease inhibitors (plasminogen activator inhibitor-1, α 2-macroglobulin and α 2-antiplasmin). Following platelet-induced haemostasis and release of TGF- β 1 and PDGF, formation of granulation tissue is facilitated by chemotaxis of neutrophils, monocytes, fibroblasts and myofibroblasts, as well as by synthesis of new extracellular matrix (ECM) and neoangiogenesis.

Neutrophil chemotaxis is stimulated by factors such as circulating complement factor 5 (C5a), leukotriene B4, kallikrein, bacterial products (if present) and numerous factors released from platelets at the site (for example, PDGF, TGF- β , platelet-activating factor and platelet factor-4 (PF-4)). Although terminally differentiated with little biosynthetic machinery, neutrophils are capable of considerable production of cytokines/chemokines necessary for effector cell recruitment, activation and response¹⁵. These phagocytic cells initiate wound healing by serving as a source of early-response pro-inflammatory cytokines such as tumour necrosis factor- α (TNF- α)⁶⁸, and interleukin (IL)-1 α and IL-1 β ⁶⁹. These cytokines mediate leukocyte adherence to the vascular endothelium, thus targeting and restricting leukocytes to areas of repair, and initiate repair by inducing expression of matrix metalloproteinases (MMPs) and keratinocyte growth factor (KGF/FGF-7) by fibroblasts⁷⁰.

In response to tissue injury, mononuclear phagocytes (that is, macrophage progenitors) migrate from the venous system to the site of tissue injury. They are guided to the site by chemotactic factors, including PF-4, TGF- β , PDGF, chemokines (monocyte chemoattractant protein-1, -2 and -3 (MCP-1/CCL2, MCP-2/CCL8 and MCP-3/CCL7), macrophage inflammatory protein-1 α and -1 β (MIP-1 α /CCL3 and MIP-1 β /CCL4), and the cytokines IL-1 β and TNF- α . Deployment of monocytes/macrophages to the site of injury peaks as the number of neutrophils decline. Once present, however, they differentiate into mature macrophages or immature dendritic cells⁷¹. After activation, macrophages are the main source of growth factors and cytokines (TGF- β 1, PDGF, bFGF, TGF- α , insulin-like growth factor (IGF)-I and -II, TNF- α and IL-1) that modulate tissue repair. Cells in their local microenvironment (for example, endothelial, epithelial, mesenchymal or neuroendocrine cells) are profoundly affected by macrophage products. Macrophages also regulate local tissue remodelling by inducing ECM components, stimulating production of proteolytic enzymes (for example, MMPs and urokinase-type plasminogen activator (uPA)), clearing apoptotic and necrotic cells, and modulating angiogenesis through local production of thrombospondin-1 (refs 72, 73).

Following their activation, mast cells are full of stored and newly synthesized inflammatory mediators. This cell type synthesizes and stores histamine, cytokines and proteases complexed to highly sulphated proteoglycans within granules, and produces lipid mediators and cytokines upon stimulation. Once activated by complement or by binding of antigens to immunoglobulin E (IgE) bound to high-affinity IgE receptors (Fc ϵ R1), they degranulate, releasing mediators including heparin, heparanase, histamine, MMPs and serine proteases, and various polypeptide growth factors, including bFGF and vascular endothelial growth factor⁷⁴. These function both in the early initiation phase of inflammation (for example, vascular reaction and exudation), and in the late phase where leukocyte accumulation and wound healing takes place.

Chemotactic cytokines

Chemokines are classified into polypeptide groups identified by the location of cysteine residues near their amino termini (for example, C-C, C-X-C, C and CX₃C). Chemokines represent the largest family of cytokines (~41 human members), forming a complex network for the chemotactic activation of all leukocytes. Chemokine receptors, members of the seven-transmembrane-spanning G-protein-coupled receptors, vary by cell type and degree of cell activation⁶. There is considerable redundancy in chemokine-receptor interaction, as many ligands bind different receptors, or vice versa.

The composition of chemokines produced at sites of tissue wounding not only recruits downstream effector cells (as discussed above), but also dictates the natural evolution of immune reactivity. For example, MCP-1/CCL2, a potent chemotactic protein for monocytes and lymphocytes, simultaneously induces expression of lymphocyte-derived IL-4 in response to antigen challenge while decreasing expression of IL-12 (ref. 75). The net effect of this alteration facilitates a switch from a T_H1-type to a T_H2-type inflammatory response.

Tissue repair

In response to wounding, fibroblasts migrate into the wound bed and initially secrete collagen type III, which is later replaced by collagen type I. Synthesis and deposition of these collagens by fibroblasts is stimulated by factors including TGF- β 1, - β 2 and - β 3, PDGF, IL-1 α , -1 β and -4, and mast cell tryptase. Once sufficient collagen has been generated, its synthesis is stopped; thus, during wound repair, production as well as the degradation of collagens is under precise spatial and temporal control.

The final phase of the healing process is re-epithelialization and migration of epithelial cells across this amalgam, in a process that requires both dissolution of the fibrin clot and degradation of the underlying dermal collagen. Epithelial cells at the leading edge of the wound express the uPA receptor, which is important for focal activation of uPA, and collagenolytic enzymes of the MMP family. In the absence of the fibrinolytic enzyme plasmin, derived from plasminogen after activation by uPA and tissue-PA, re-epithelialization is dramatically delayed⁷⁶.

The pro-inflammatory properties of TGF- β , such as leukocyte recruitment, adhesion and regulation of MMP secretion and activation, are balanced by its ability to also reverse its role, and suppress these events and foster ECM synthesis to mediate tissue repair⁸. As inflammatory cells are activated, their complement of TGF- β receptors change, resulting in differential susceptibility to TGF- β and enhanced sensitivity to suppression by TGF- β ⁸, a critical event to resolving inflammation.

T cells. During development of melanoma, activated macrophages produce TGF- β , TNF- α , IL-1 α , arachidonate metabolites and extracellular proteases¹⁸. In response, melanocytes express IL-8 and vascular endothelial growth factor (VEGF)-A, thereby inducing vascular angiogenesis under paracrine control¹⁸. Indeed, macrophage

infiltration is closely associated with the depth of invasion of primary melanoma due, in part, to macrophage-regulated tumour-associated angiogenesis¹⁹.

In addition to altering the local balance of pro-angiogenic factors during melanoma development, during human cervical carcinogen-

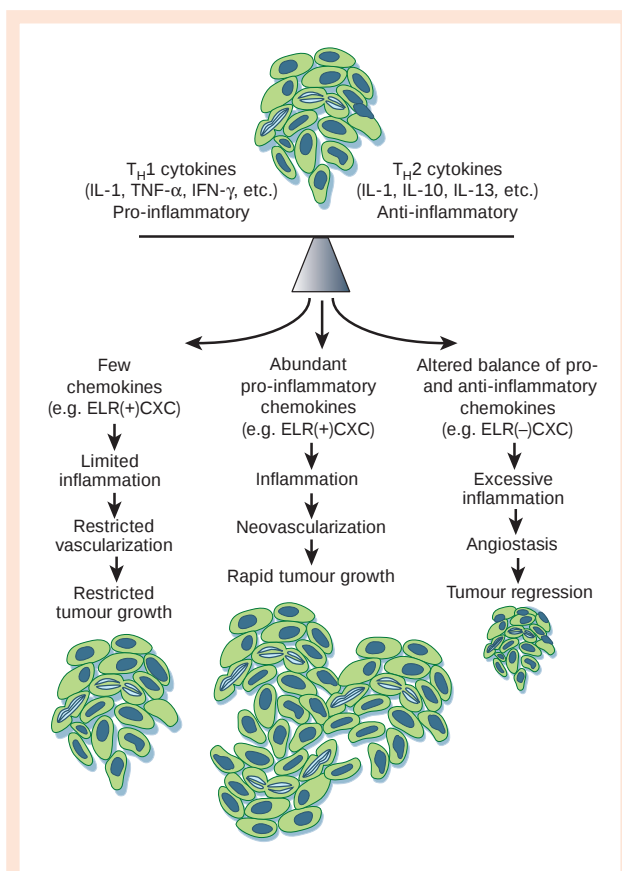


Figure 2 Cytokine and chemokine balances regulate neoplastic outcome. The balance of cytokines in any given tumour is critical for regulating the type and extent of inflammatory infiltrate that forms. Tumours that produce little or no cytokines or an overabundance of anti-inflammatory cytokines induce limited inflammatory and vascular responses, resulting in constrained tumour growth. In contrast, production of an abundance of pro-inflammatory cytokines can lead to a level of inflammation that potentiates angiogenesis, thus favouring neoplastic growth. Alternatively, high levels of monocytes and/or neutrophil infiltration, in response to an altered balance of pro-versus anti-inflammatory cytokines, can be associated with cytotoxicity, angiostasis and tumour regression. In tumours, interleukin-10 is generally a product of tumour cells and tumour-associated macrophages.

esis, TAMs express VEGF-C and VEGF-D as well as the VEGF receptor-3 (VEGFR-3), all of which are implicated in formation of lymphatic vessels and lymphatic metastases¹⁷. By placing TAMs at the centre of the recruitment and response to angiogenic and lymphangiogenic stimuli, they may foster the spread of tumours. TAMs also induce VCAM-1 expression on mesothelial cells, a step also believed to be key for tumour cell dissemination into the peritoneum²⁰.

The functional significance of macrophage recruitment to sites of neoplastic growth has been examined by crossing transgenic mice expressing Polyoma virus middle T (PyMT) driven by the mouse mammary tumour virus (MMTV) long terminal repeat, which are prone to development of mammary cancer, with mice containing a null mutation in the CSF-1 gene (*Csf1^{op}*)²¹. Whereas the absence of CSF-1 during early neoplastic development is without apparent consequence, development of late-stage invasive carcinoma and pulmonary metastases are significantly attenuated. The key difference between PyMT mice and PyMT/*Csf1^{op}*/*Csf1^{op}* mice is not in the apparent proliferative capacity of neoplastic epithelial cells, but in the failure to recruit mature macrophages into neoplastic tissue in the absence of CSF-1. Targeting CSF-1 expression specifically to

mammary epithelium in CSF-1-null/PyMT mice restores macrophage recruitment, primary tumour development and metastatic potential¹². A similar study showed that subcutaneous growth of Lewis lung cancer cells is impaired in *Csf1^{op}*/*Csf1^{op}* mice²². In this example, however, tumours displayed a decreased mitotic index and pronounced necrosis, apparently resulting from diminished angiogenesis and impaired tumour-stroma formation. These defects were corrected by treatment of tumour-bearing mice with recombinant CSF-1 (ref. 22). Together, these genetic experiments provide a causal link between CSF-1-dependent infiltrating macrophages and the malignant potential of epithelial cells.

Macrophages are not unique among inflammatory cells in potentiation of neoplastic processes. Genetic and functional experiments indicate that neutrophils, mast cells, eosinophils and activated T lymphocytes also contribute to malignancies by releasing extracellular proteases, pro-angiogenic factors and chemokines^{11,23–26}.

Cancers associated with chronic inflammation

How are inflammatory cells co-opted into the neoplastic process? A plausible hypothesis is that many malignancies arise from areas of infection and inflammation, simply as part of the normal host response. Indeed, there is a growing body of evidence that many malignancies are initiated by infections^{11,27–29} (Table 1) — upwards of 15% of malignancies worldwide can be attributed to infections, a global total of 1.2 million cases per year¹¹. Persistent infections within the host induce chronic inflammation. Leukocytes and other phagocytic cells induce DNA damage in proliferating cells, through their generation of reactive oxygen and nitrogen species that are produced normally by these cells to fight infection³⁰. These species react to form peroxynitrite, a mutagenic agent³⁰. Hence, repeated tissue damage and regeneration of tissue, in the presence of highly reactive nitrogen and oxygen species released from inflammatory cells, interacts with DNA in proliferating epithelium resulting in permanent genomic alterations such as point mutations, deletions, or rearrangements. Indeed, p53 mutations are seen at frequencies similar to those in tumours in chronic inflammatory diseases such as rheumatoid arthritis and inflammatory bowel disease³¹.

The strongest association of chronic inflammation with malignant diseases is in colon carcinogenesis arising in individuals with inflammatory bowel diseases, for example, chronic ulcerative colitis and Crohn's disease. Hepatitis C infection in the liver predisposes to liver carcinoma, an increased risk of bladder and colon carcinoma is associated with schistosomiasis, whereas chronic *Helicobacter pylori* infection is the world's leading cause of stomach cancer³². The Gram-negative bacterium *H. pylori* is established as a definite carcinogen for the development of gastric cancer — the second most common type of cancer globally^{11,29} — and DNA damage resulting from chronic inflammation is believed the mechanism³². Exacerbating DNA damage induced by inflammatory cells is expression of macrophage migration inhibitory factor (MIF) from macrophages and T lymphocytes. MIF is a potent cytokine that overcomes p53 function by suppressing its transcriptional activity³³. Chronic bypass of p53 regulatory functions in infiltrated tissues can enhance proliferation and extend life span, while also creating an environment with a deficient response to DNA damage, amplifying accumulation of potential oncogenic mutations.

Infectious viral agents, for example, DNA tumour viruses, may also directly transform cells by inserting active oncogenes into the host genome, although other mechanisms also are responsible. While many types of infectious agents are present in animals, only a subset of individuals infected with human papilloma virus, hepatitis B virus (HBV) or Epstein-Barr virus develop virus-associated malignancies. This may reflect immune suppression, the necessity of cofactors necessary for promotion or the fact that a neoplasm can develop only if viral infection has targeted a pluripotent progenitor or stem cell. Such stem cells are typically low in abundance and located in regions of tissues protected from agents that would otherwise

harm them³⁴. In Rous sarcoma virus infections, inflammation is essential for tumour development and this requirement is mediated by factors such as TGF- β and other cytokines produced by the inflammatory cells³⁵. Epstein-Barr virus also causes sustained proliferation of B lymphocytes, which, when coupled with a secondary mutation, can result in neoplastic progression and malignant conversion to give rise to Burkett's lymphoma.

The molecular mechanism behind the associated risk of hepatocellular carcinoma resulting from HBV and/or hepatitis C virus (HCV) infection is uncertain. Although there is evidence for clonal integration of viral DNA in tumours and surrounding parenchyma cells, there are no defined transforming sequences found within the viral genomes that can act as viral oncogenes. Moreover, there is no evidence to suggest that viral integration activates either a classical cellular oncogene or inactivates a cellular tumour suppressor gene. HCV core protein interacts with the signal transducer and activator of transcription 3 (STAT3) protein³⁶, a transcription factor involved in mediating cytokine signalling³⁷. This interaction induces sustained phosphorylation of a critical tyrosine residue, resulting in enhanced proliferation and upregulation of Bcl-x_L and cyclin-D. Thus, chronic viral replication in hepatocytes may alter the local cytokine profile and the apoptotic or proliferative responses in infected cells, with an immune response to the viral proteins resulting in a state of chronic inflammation. Interestingly, a similar pathway involving inflammation, IL-6 and STAT3 is downstream of *H. pylori* in the generation of stomach cancer³⁸.

The chemokine connection

Chemokines were initially defined functionally as soluble factors regulating directional migration of leukocytes during states of inflammation; however, chemokine biology extends to all cell types, including most human neoplastic cells⁶. Attention first focused on the role of chemokines during malignancy when it was reported that experimental animals without T or natural killer (NK) cell functions, when challenged with a tumour, showed a typical inflammatory infiltrate; this suggested that neoplastic cells either produce chemotactic factors or induce their expression in nearby 'host' cells³⁹. It is now appreciated that the chemokine-receptor system can be altered dramatically in neoplastic tissue, particularly at the invasive edges. Moreover, chemokines induce direct effects on stromal and neoplastic cells in addition to their roles in regulating leukocyte recruitment (Fig. 2).

Regulation of tumour growth by chemokines. Some tumour cells not only regulate their chemokine expression to help recruit inflammatory cells, but also use these factors to further the tumour growth and progression. Melanoma is perhaps the best exemplar in which chemokines (for example, GRO α /CXCL1, GRO β /CXCL2, GRO γ /CXCL3 and IL-8/CXCL8) have been shown to exert autocrine control over neoplastic cell proliferation⁴⁰. Blocking GRO α or the CXCR2 receptor attenuates melanoma cell proliferation *in vitro*⁴¹, whereas overexpression of GRO α , GRO β or GRO γ in a variety of tumour-derived cell lines enhances their colony-forming activity and tumorigenicity in nude mice^{42,43}. Other CXCR2 ligands have been identified as having autocrine roles in the growth of pancreatic, head and neck, and non-small-cell lung carcinoma^{44,45}, whereas in mouse models, ENA-78/CXCL5 variably affects tumour growth, vascularity and apoptosis⁴⁶. Macrophage pro-inflammatory chemokine-3 α (MIP-3 α /CCL20), a CC chemokine, is overexpressed in pancreatic carcinoma cells and infiltrating macrophages adjacent to tumours; MIP-3 α /CCL20 stimulates growth of neoplastic cells while simultaneously enhancing migration of TAMs⁴⁷.

Regulation of angiogenesis by chemokines. Activation of angiogenic programmes represents a shift in the balance between pro- and anti-angiogenic factors⁴⁸. Although angiogenesis is strictly controlled, it is associated with chronic inflammatory diseases, such as psoriasis, rheumatoid arthritis and fibrosis, as well as with tumour growth and metastasis⁴⁸. It is well established that CXC chemokines with the three amino acids (Glu-Leu-Arg/ELR) immediately amino-terminal

Table 1 Chronic inflammatory conditions associated with neoplasms

Pathologic condition	Associated neoplasm(s)	Aetiological agent
Asbestosis, silicosis	Mesothelioma, lung carcinoma	Asbestos fibres, silica particles
Bronchitis	Lung carcinoma	Silica, asbestos, smoking (nitrosamines, peroxides)
Cystitis, bladder inflammation	Bladder carcinoma	Chronic indwelling, urinary catheters
Gingivitis, lichen planus	Oral squamous cell carcinoma	
Inflammatory bowel disease, Crohn's disease, chronic ulcerative colitis	Colorectal carcinoma	
Lichen sclerosis	Vulvar squamous cell carcinoma	
Chronic pancreatitis, hereditary pancreatitis	Pancreatic carcinoma	Alcoholism, mutation in trypsinogen gene on Ch. 7
Reflux oesophagitis, Barrett's oesophagus	Oesophageal carcinoma	Gastric acids
Sialadenitis	Salivary gland carcinoma	
Sjögren syndrome, Hashimoto's thyroiditis	MALT lymphoma	
Skin inflammation	Melanoma	Ultraviolet light
Cancers associated with infectious agents		
<i>Opisthorchis</i> , <i>Cholangitis</i>	Cholangiosarcoma, colon carcinoma	Liver flukes (<i>Opisthorchis viverrini</i>), bile acids
Chronic cholecystitis	Gall bladder cancer	Bacteria, gall bladder stones
Gastritis/ulcers	Gastric adenocarcinoma, MALT	<i>Helicobacter pylori</i>
Hepatitis	Hepatocellular carcinoma	Hepatitis B and/or C virus
Mononucleosis	B-cell non-Hodgkin's lymphoma, Burkitt's lymphoma	Epstein-Barr Virus
AIDS	Non-Hodgkin's lymphoma, squamous cell carcinomas, Kaposi's sarcoma	Human immunodeficiency virus, human herpesvirus type 8
Osteomyelitis	Skin carcinoma in draining sinuses	Bacterial infection
Pelvic inflammatory disease, chronic cervicitis	Ovarian carcinoma, cervical/anal carcinoma	Gonorrhoea, chlamydia, human papillomavirus
Chronic cystitis	Bladder, liver, rectal carcinoma, follicular lymphoma of the spleen	Schistosomiasis

Modified from refs 29, 67. MALT, mucosa-associated lymphoid tissue.

to the CXC motif (ELR⁺) are pro-angiogenic and stimulate endothelial cell chemotaxis, whereas ELR⁻ CXC chemokines (for example, PF-4/CXCL4, MIG/CXCL9 and IP-10/CXCL10) possess angiostatic activities^{44,49}. ELR⁺ CXC ligands bind to CXCR2 and to a lesser degree to CXCR1, whereas ELR⁻ CXC ligands bind to CXCR3, CXCR4 and CXCR5 (ref. 6). Compared to VEGF-A, murine MCP-5/CCL12 exhibits only modest mitogenic properties towards endothelial cells; however, it is a potent chemoattractant. In contrast, stromal-cell-derived factor 1 (SDF-1/CXCL12) induces endothelial expression of VEGF-A; VEGF-A in turn upregulates CXCR4 on endothelial cells⁷. Although it is not always clear if the angiostatic and angiogenic effects of chemokines are direct or indirect, it is accepted that the balance between the two regulates neoplastic cell physiology.

Chemokines and metastasis. Malignant cells that possess metastatic capacity have properties endowing them with the ability to invade and survive in ectopic tissue, venous and/or lymphatic environments, as well as ability to reside and proliferate at a distal site (Fig. 3). Much debate exists as to whether malignant cells metastasize to environments favouring their specific growth or whether different organs are endowed with the ability to arrest or attract specific types of malignant cells through chemotactic factors (the so-called homing theory)⁴⁸. Studies using a mouse model by Muller and colleagues suggest that the pattern of breast cancer metastases is in part governed by specific interactions between CXCR4 and its ligand SDF-1/CXCL12 (ref. 50). CXCL12 is a rather unique chemokine in that it is the product of resting cells in multiple organs⁵, and is particu-

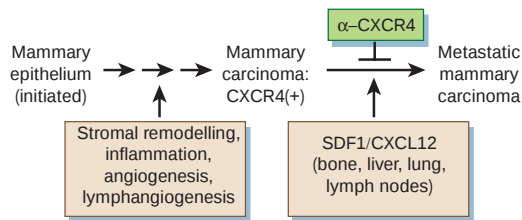


Figure 3 Cancer metastasis and chemokine signalling. Initiated epithelial cells are promoted by inflammation to undergo neoplastic progression, a process that requires remodelling of the extracellular matrix, recruitment of inflammatory cells, angiogenesis and lymphangiogenesis. Out of this microenvironment, carcinomas arise. These neoplastic cells then turn on expression of chemokine receptors, such as CXCR4. The production of chemokine ligands for these receptors, in sites such as lymph nodes, bone marrow, liver and lung, then facilitates their invasion and migration to secondary sites where malignant cells reside either in a dormant state, or proliferate to form a productive metastatic lesion. Blockade of chemokine receptors, for example, anti-CXCR4 antibodies, attenuates metastatic spread in some experimental systems.

larly highly expressed in target organs for breast cancer metastasis⁵⁰. CXCL12 triggers chemotaxis of malignant mammary carcinoma cells *in vitro*, and the chemotactic activity of extracts of organs targeted by breast cancer cells (bone marrow, liver, lung and lymph nodes) can be neutralized by anti-CXCR4 antibodies. The involvement of CXCR4 in metastasis is not limited to breast cancer, as CXCR4 is expressed in tumour cell lines (for example, prostate carcinomas, B-cell lymphomas, astroglomas and chronic lymphocytic leukaemias) that also respond to CXCL12 (ref. 51). The broader implications of these observations are that chemokines may be involved in regulating the spectrum of metastases in diverse cancer types.

Tumours commandeer leukocyte adhesion mechanisms

Tumour cells not only take advantage of the trophic factors made by inflammatory cells, but may also use the same adhesion molecules, chemokines and receptors to aid in migration and homing during distant metastatic spread. Evidence suggests that mechanisms used for homing of leukocytes may be appropriated for the dissemination of tumours via the bloodstream and lymphatics. Selectins are adhesion receptors that normally recognize certain vascular mucin-type glycoproteins bearing the carbohydrate structure sialyl-Lewis X and facilitate leukocyte rolling along the blood vessels. Metastatic progression of many epithelial carcinomas correlates with tumour production of mucins containing sialyl-Lewis X. Lung colonization by melanoma cells that express sialyl-Lewis X is significantly reduced in E/P-selectin-deficient mice⁵². P-selectin deficiency attenuates tumour growth and metastasis, and tumours are significantly smaller in mice treated with a receptor antagonist peptide.

These results indicate that receptors expressed in the vasculature are crucial in targeting sialyl-Lewis X-dependent cancer cells⁵³. P-selectin facilitates human carcinoma metastasis in immunodeficient mice by mediating early interactions of platelets with blood-borne tumour cells via their cell-surface mucins, a process that can be blocked by heparin⁵⁴. L-selectin on neutrophils, monocytes and/or NK cells also may facilitate metastasis⁵⁵. Metastasis could involve the formation of tumour-platelet-leukocyte emboli that interact with the vasculature of distant organs. In addition, expression of L-selectin on tumour cells can foster metastasis to lymph nodes⁵⁶.

Inflammation as an anti-cancer therapeutic opportunity

Perhaps the best evidence for the significance of inflammation during neoplastic progression comes from study of cancer risk among

long-term users of aspirin and nonsteroidal anti-inflammatory drugs (NSAIDs). Much data indicates that use of these drugs reduces colon cancer risk by 40–50%, and may be preventative for lung, oesophagus and stomach cancer^{57,58}. The ability of NSAIDs to inhibit cyclo-oxygenases (COX-1 and -2) underlies their mechanism(s) of chemoprevention. COX-2 converts arachidonic acid to prostaglandins, which in turn induces inflammatory reactions in damaged tissues⁵⁹. Aspirin is non-selective in its inhibition of platelet function by acetylating and irreversibly inactivating both COX-1 and COX-2. Inactivation prevents platelet synthesis of prostaglandins, endoperoxides and thromboxane A₂.

Other NSAIDs, for example, flurbiprofen, may have strong anti-metastatic effects because of their inhibition of platelet aggregation⁶⁰. But NSAIDs may act through mechanisms other than inhibition of COX enzyme activity alone, as some NSAIDs lacking COX-inhibitory function show efficacy in inhibiting colon carcinogenesis⁶¹. Other mechanisms have been proposed¹⁵, including induction of apoptosis through release of cytochrome C from mitochondria and subsequent activation of caspase-9 and -3, and/or interference with cell-cycle progression, reduction of carcinogen activation and stimulation of immune surveillance.

The pro-inflammatory cytokine TNF- α is also a key downstream mediator in inflammation. Despite the name, TNF- α is important in early events in tumours, regulating a cascade of cytokines, chemokines, adhesions, MMPs and pro-angiogenic activities^{1,62}. Thus, TNF- α may be one of the ways in which inflammation acts as a tumour promoter. Blocking antibodies that have significant therapeutic efficacy in other inflammatory diseases⁶³ may have applications in therapy in cancer.

Tumours are also rich in mucins and other ligands that may include the sialyl-Lewis X epitope recognized by selectins. Because selectins may have a role in metastasis^{54,55}, targeting the selectin interaction with heparin or antagonists of the receptor may decrease metastasis⁵⁴.

MMPs are produced by inflammatory cells and by stromal cells responding to chemokines and cytokines produced by inflammatory cells in tumour microenvironments²⁵. Like inflammatory cells, MMPs may both promote tumour progression and attenuate it. Indeed, MMPs may mediate many of the actions of inflammatory cells in neoplasms⁶⁴. MMPs can recruit inflammatory cells by releasing chemoattractants and motogens; they also generate growth-promoting and cytostatic signals. MMPs activate angiogenesis, but also produce fragments of basement-membrane collagens and plasminogen that are angiogenesis inhibitors. They have both apoptotic and anti-apoptotic actions. Thus, the efficacy of MMP inhibitors may be mediated, at least in part, through anti-inflammatory actions^{64,65}. Given their diverse actions, it is also not surprising that trials with MMP inhibitors have had mixed results, with efficacy reported mostly during early tumour progression⁶⁶.

Inflammatory cells and cancer: friend or foe?

It is now evident that inflammatory cells have powerful effects on tumour development. Early in the neoplastic process, these cells are powerful tumour promoters, producing an attractive environment for tumour growth, facilitating genomic instability and promoting angiogenesis. The inflammatory cells, and the chemokines and cytokines that they produce, influence the whole tumour organ, regulating the growth, migration and differentiation of all cell types in the tumour microenvironment, including neoplastic cells, fibroblasts and endothelial cells. Later in the tumorigenic process, neoplastic cells also divert inflammatory mechanisms such as selectin-ligand interactions, MMP production and chemokine functions to favour neoplastic spread and metastasis. This may be part of an attempt by the tumour to subvert immune cell functions, so favouring tumour development. Yet, the recruitment of inflammatory cells may also be counterproductive for tumour development, and also may represent an attempt by the host to suppress tumour growth.

The pro-tumour actions of inflammatory cells include releasing growth and survival factors, promoting angiogenesis and lymphangiogenesis, stimulating DNA damage, remodelling the ECM to facilitate invasion, coating tumour cells to make available receptors for disseminating cells via lymphatics and capillaries, and evading host defence mechanisms. Although inflammatory responses should also be anti-tumour, cancer patients are often defective in their inflammatory responses. This may arise by two distinct tumour-mediated mechanisms: a failure to upregulate the anti-inflammatory cytokines, or subversion of the host response resulting from desensitization of receptors owing to high chemokine and cytokine concentrations that then blunt systemic responses. Can we apply these new insights for targeting metastases?

It is clear that anti-inflammatory therapy is efficacious towards early neoplastic progression and malignant conversion. In a fully developed malignancy, there are 'excess' inflammatory cells in the tumour microenvironment. Does the tumour need inflammation to help foster angiogenesis? We must think globally and act locally. One approach is to evaluate whether functional polymorphisms in genes that regulate inflammatory processes (for example, genes encoding MMPs, cytokines, chemokines or selectins) harbour altered risk for developing cancer or are indicators of prognosis. Yet for all the local inflammation in tumours, in many cases the overall innate immunity of the host is blunted. The challenge for the future is to normalize the inflammatory network to regain a normal host response overall: decreasing the high levels of tumour-promoting properties of the infiltrating cells, such as pro-inflammatory cytokines, while increasing their tumour-suppressing properties, such as anti-inflammatory cytokines. In this way, later in tumour progression, we can harness the activities that are anti-tumour while suppressing those that are pro-tumour. □

doi:10.1038/nature01322

1. Balkwill, F. & Mantovani, A. Inflammation and cancer: back to Virchow? *Lancet* **357**, 539–545 (2001).
2. Dvorak, H. F. Tumors: wounds that do not heal. Similarities between tumor stroma generation and wound healing. *N. Engl. J. Med.* **315**, 1650–1659 (1986).
3. Dranoff, G. Tumour immunology: immune recognition and tumor protection. *Curr. Opin. Immunol.* **14**, 161–164 (2002).
4. Pardoll, D. M. Spinning molecular immunology into successful immunotherapy. *Nature Rev. Immunol.* **2**, 227–238 (2002).
5. Chetibi, S. & Ferguson, M. W. J. In *Inflammation: Basic Principles and Clinical Correlates* (eds Gallin, J. I. & Snyderman, R.) 865–881 (Lippincott, Williams and Wilkins, Philadelphia, 1999).
6. Rossi, D. & Zlotnik, A. The biology of chemokines and their receptors. *Annu. Rev. Immunol.* **18**, 217–242 (2000).
7. Homey, B., Muller, A. & Zlotnik, A. Chemokines: agents for the immunotherapy of cancer? *Nature Rev. Immunol.* **2**, 175–184 (2002).
8. Moustakas, A., Pardali, K., Gaal, A. & Heldin, C. H. Mechanisms of TGF- β signaling in regulation of cell growth and differentiation. *Immunol. Lett.* **82**, 85–91 (2002).
9. Rous, P. & Kidd, J. Conditional neoplasms and subthreshold neoplastic states: a study of the tar tumors of rabbits. *J. Exp. Med.* **73**, 365–389 (1941).
10. Mackenzie, I. C. & Rous, P. The experimental disclosure of latent neoplastic changes in tarred skin. *J. Exp. Med.* **73**, 391–415 (1941).
11. Kuper, H., Adami, H. O. & Trichopoulos, D. Infections as a major preventable cause of human cancer. *J. Intern. Med.* **248**, 171–183 (2000).
12. Wahl, L. M. & Kleinman, H. K. Tumor-associated macrophages as targets for cancer therapy. *J. Natl Cancer Inst.* **90**, 1583–1584 (1998).
13. Talmor, M. et al. Generation of large numbers of immature and mature dendritic cells from rat bone marrow cultures. *Eur. J. Immunol.* **28**, 811–817 (1998).
14. Allavena, P. et al. The chemokine receptor switch paradigm and dendritic cell migration: its significance in tumor tissues. *Immunol. Rev.* **177**, 141–149 (2000).
15. Brigati, C., Noonan, D. M., Albini, A. & Benelli, R. Tumors and inflammatory infiltrates: friends or foes? *Clin. Exp. Metastasis* **19**, 247–258 (2002).
16. Tsung, K., Dolan, J. P., Tsung, Y. L. & Norton, J. A. Macrophages as effector cells in interleukin 12-induced T cell-dependent tumor rejection. *Cancer Res.* **62**, 5069–5075 (2002).
17. Schoppmann, S. et al. Tumor-associated macrophages express lymphatic endothelial growth factors and are related to peritumoral lymphangiogenesis. *Am. J. Pathol.* **161**, 947–956 (2002).
18. Torisu, H. et al. Macrophage infiltration correlates with tumor stage and angiogenesis in human malignant melanoma: possible involvement of TNF α and IL-1 α . *Int. J. Cancer* **85**, 182–188 (2000).
19. Ono, M., Torisu, H., Fukushima, J., Nishie, A. & Kuwano, M. Biological implications of macrophage infiltration in human tumor angiogenesis. *Cancer Chemother. Pharmacol.* **43**, 569–571 (1999).
20. Jonjic, N. et al. Expression of adhesion molecules and chemotactic cytokines in cultured human mesothelial cells. *J. Exp. Med.* **176**, 1165–1174 (1992).
21. Lin, E. Y., Nguyen, A. V., Russell, R. G. & Pollard, J. W. Colony-stimulating factor 1 promotes progression of mammary tumors to malignancy. *J. Exp. Med.* **193**, 727–740 (2001).
22. Nowicki, A. et al. Impaired tumor growth in colony-stimulating factor 1 (CSF-1)-deficient, macrophage-deficient op/op mouse: evidence for a role of CSF-1-dependent macrophages in formation of tumor stroma. *Int. J. Cancer* **65**, 112–119 (1996).

23. DiCarlo, E. et al. The intriguing role of polymorphonuclear neutrophils in antitumor reactions. *Blood* **97**, 339–345 (2001).
24. Coussens, L. M. et al. Inflammatory mast cells up-regulate angiogenesis during squamous epithelial carcinogenesis. *Genes Dev.* **13**, 1382–1397 (1999).
25. Coussens, L. M., Tinkle, C. L., Hanahan, D. & Werb, Z. MMP-9 supplied by bone marrow-derived cells contributes to skin carcinogenesis. *Cell* **103**, 481–490 (2000).
26. Bergers, G. et al. Matrix metalloproteinase-9 triggers the angiogenic switch during carcinogenesis. *Nature Cell Biol.* **2**, 737–744 (2000).
27. Blaser, M. J., Chyou, P. H. & Nomura, A. Age at establishment of *Helicobacter pylori* infection and gastric carcinoma, gastric ulcer, and duodenal ulcer risk. *Cancer Res.* **55**, 562–565 (1995).
28. Scholl, S. M. et al. Anti-colony-stimulating factor-1 antibody staining in primary breast adenocarcinomas correlates with marked inflammatory cell infiltrates and prognosis. *J. Natl Cancer Inst.* **86**, 120–126 (1994).
29. Shacter, E. & Weitzman, S. A. Chronic inflammation and cancer. *Oncology* **16**, 217–226 (2002).
30. Maeda, H. & Akaike, T. Nitric oxide and oxygen radicals in infection, inflammation, and cancer. *Biochemistry* **63**, 854–865 (1998).
31. Yamanishi, Y. et al. Regional analysis of p53 mutations in rheumatoid arthritis synovium. *Proc. Natl Acad. Sci. USA* **99**, 10025–10030 (2002).
32. Ernst, P. B. & Gold, B. D. The disease spectrum of *Helicobacter pylori*: the immunopathogenesis of gastroduodenal ulcer and gastric cancer. *Annu. Rev. Microbiol.* **54**, 615–640 (2000).
33. Hudson, J. D. et al. A proinflammatory cytokine inhibits p53 tumor suppressor activity. *J. Exp. Med.* **190**, 1375–1382 (1999).
34. Jensen, U. B., Lowell, S. & Watt, F. M. The spatial relationship between stem cells and their progeny in the basal layer of human epidermis: a new view based on whole-mount labeling and lineage analysis. *Development* **126**, 2409–2418 (1999).
35. Martins-Green, M., Boudreau, N. & Bissell, M. J. Inflammation is responsible for the development of wound-induced tumors in chickens infected with Rous sarcoma virus. *Cancer Res.* **54**, 4334–4341 (1994).
36. Yoshida, T. et al. Activation of STAT3 by the hepatitis C virus core protein leads to cellular transformation. *J. Exp. Med.* **196**, 641–653 (2002).
37. Bromberg, J. & Darnell, J. E. The role of STATs in transcriptional control and their impact on cellular function. *Oncogene* **19**, 2468–2473 (2000).
38. Tebbutt, N. C. et al. Reciprocal regulation of gastrointestinal homeostasis by SHP2 and STAT-mediated trefoil gene activation in gp130 mutant mice. *Nature Med.* **8**, 1089–1097 (2002).
39. Mantovani, A., Muzio, M., Garlanda, C., Sozzani, S. & Allavena, P. Macrophage control of inflammation: negative pathways of regulation of inflammatory cytokines. *Novartis Found. Symp.* **234**, 120–131 (2001).
40. Richmond, A. & Thomas, H. Purification of melanoma growth stimulatory activity. *J. Cell. Physiol.* **129**, 375–384 (1986).
41. Norgauer, J., Metzner, B. & Schraufstatter, I. Expression and growth-promoting function of the IL-8 receptor β in human melanoma cells. *J. Immunol.* **156**, 1132–1137 (1996).
42. Balentien, E., Mufson, B. E., Shattuck, R. L., Derynck, R. & Richmond, A. Effects of MGSA/GRO alpha on melanocyte transformation. *Oncogene* **6**, 1115–1124 (1991).
43. Owen, J. D. et al. Enhanced tumor-forming capacity for immortalized melanocytes expressing melanoma growth stimulatory activity/growth-regulated cytokine beta and gamma proteins. *Int. J. Cancer* **73**, 94–103 (1997).
44. Vicari, A. P. & Caux, C. Chemokines in cancer. *Cytokine Growth Factor Rev.* **13**, 143–154 (2002).
45. Farrow, B. & Evers, B. M. Inflammation and the development of pancreatic cancer. *Surg. Oncol.* **10**, 153–169 (2002).
46. Arenberg, D. A. et al. Epithelial-neutrophil activating peptide (ENA-78) is an important angiogenic factor in non-small cell lung cancer. *J. Clin. Invest.* **102**, 465–472 (1998).
47. Kleeff, J. et al. Detection and localization of Mip-3 α /LARC/Exodus, a macrophage proinflammatory chemokine, and its CCR6 receptor in human pancreatic cancer. *Int. J. Cancer* **81**, 650–657 (1999).
48. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**, 57–70 (2000).
49. Strieter, R. M. et al. The functional role of the ELR motif in CXC chemokine-mediated angiogenesis. *J. Biol. Chem.* **270**, 27348–27357 (1995).
50. Muller, A. et al. Involvement of chemokine receptors in breast cancer metastasis. *Nature* **410**, 50–56 (2001).
51. Moore, M. A. The role of chemoattraction in cancer metastases. *BioEssays* **23**, 674–676 (2001).
52. Kim, Y. J., Borsig, L., Varki, N. M. & Varki, A. P-selectin deficiency attenuates tumor growth and metastasis. *Proc. Natl Acad. Sci. USA* **95**, 9325–9330 (1998).
53. Zhang, J. et al. Sialyl Lewis X-dependent lung colonization of B16 melanoma cells through a selectin-like endothelial receptor distinct from E- or P-selectin. *Cancer Res.* **62**, 4194–4198 (2002).
54. Borsig, L. et al. Heparin and cancer revisited: mechanistic connections involving platelets, P-selectin, carcinoma mucins, and tumor metastasis. *Proc. Natl Acad. Sci. USA* **98**, 3352–3357 (2001).
55. Borsig, L., Wong, R., Hynes, R. O., Varki, N. M. & Varki, A. Synergistic effects of L- and P-selectin in facilitating tumor metastasis can involve non-mucin ligands and implicate leukocytes as enhancers of metastasis. *Proc. Natl Acad. Sci. USA* **99**, 2193–2198 (2002).
56. Qian, F., Hanahan, D. & Weissman, I. L. L-selectin can facilitate metastasis to lymph nodes in a transgenic mouse model of carcinogenesis. *Proc. Natl Acad. Sci. USA* **98**, 3976–3981 (2001).
57. Baron, J. A. & Sandler, R. S. Nonsteroidal anti-inflammatory drugs and cancer prevention. *Annu. Rev. Med.* **51**, 511–523 (2000).
58. Garcia-Rodriguez, L. A. & Huerta-Alvarez, C. Reduced risk of colorectal cancer among long-term users of aspirin and nonaspirin nonsteroidal antiinflammatory drugs. *Epidemiology* **12**, 88–93 (2001).
59. Williams, C. S., Mann, M. & DuBois, R. N. The role of cyclooxygenases in inflammation, cancer, and development. *Oncogene* **18**, 7908–7916 (1999).
60. Mamytebkova, A., Rezabek, K., Kacerovska, H., Grimova, J. & Svobodova, J. Antimetastatic effect of flurbiprofen and other platelet aggregation inhibitors. *Neoplasia* **33**, 417–421 (1986).
61. Elder, D. J., Halton, D. E., Hague, A. & Paraskeva, C. Induction of apoptotic cell death in human colorectal carcinoma cell lines by a cyclooxygenase-2 (COX-2)-selective nonsteroidal anti-inflammatory drug: independence from COX-2 protein expression. *Clin. Cancer Res.* **3**, 1679–1683 (1997).
62. Balkwill, F. Tumor necrosis factor or tumor promoting factor? *Cytokine Growth Factor Rev.* **13**, 135–141 (2002).
63. Shanahan, J. C. & St. Clair, E. W. Tumor necrosis factor- α blockade: a novel therapy for rheumatic disease. *Clin. Immunol.* **103**, 231–242 (2002).

64. Egeblad, M. & Werb, Z. New functions for the matrix metalloproteinases in cancer progression. *Nature Rev. Cancer* **2**, 161–174 (2002).
65. Overall, C. M. & Lopez-Otin, C. Strategies for MMP inhibition in cancer: innovations for the post-trial era. *Nature Rev. Cancer* **2**, 657–672 (2002).
66. Coussens, L. M., Fingleton, B. & Matrisian, L. M. Matrix metalloproteinase inhibitors and cancer: trials and tribulations. *Science* **295**, 2387–2392 (2002).
67. Dalglish, A. G. & O'Byrne, K. J. Chronic immune activation and inflammation in the pathogenesis of AIDS and cancer. *Adv. Cancer Res.* **84**, 231–276 (2002).
68. Feiken, E., Romer, J., Eriksen, J. & Lund, L. R. Neutrophils express tumor necrosis factor- α during mouse skin wound healing. *J. Invest. Dermatol.* **105**, 120–123 (1995).
69. Hubner, G. *et al.* Differential regulation of pro-inflammatory cytokines during wound healing in normal and glucocorticoid-treated mice. *Cytokine* **8**, 548–556 (1996).
70. Chedid, M., Rubin, J. S., Csaky, K. G. & Aaronson, S. A. Regulation of keratinocyte growth factor gene expression by interleukin 1. *J. Biol. Chem.* **269**, 10753–10757 (1994).
71. Osusky, R., Malik, P. & Ryan, S. J. Retinal pigment epithelium cells promote the maturation of monocytes to macrophages in vitro. *Ophthalmic Res.* **29**, 31–36 (1997).
72. DiPietro, L. Wound healing: the role of the macrophage and other immune cells. *Shock* **4**, 233–240 (1995).
73. Fritsch, C., Simon-Assmann, P., Kedinger, M. & Evans, G. S. Cytokines modulate fibroblast phenotype and epithelial-stroma interactions in rat intestine. *Gastroenterology* **112**, 826–838 (1997).
74. Grutzkau, A. *et al.* Synthesis, storage, and release of vascular endothelial growth factor/vascular permeability factor (VEGF/VPF) by human mast cells: implications for the biological significance of VEGF206. *Mol. Biol. Cell* **9**, 875–884 (1998).
75. Chensue, S. W., Ruth, J. H., Warmington, K., Lincoln, P. & Kunkel, S. L. In vivo regulation of macrophage IL-12 production during type 1 and type 2 cytokine-mediated granuloma formation. *J. Immunol.* **155**, 3546–3551 (1995).
76. Romer, J. *et al.* Impaired wound healing in mice with a disrupted plasminogen gene. *Nature Med.* **2**, 287–292 (1996).

Acknowledgements

Supported by grants from the National Institutes of Health, the American Cancer Society, the V Foundation for Cancer Research, the Edward Mallinckrodt Jr Foundation for Medical Research, and the American Association for Cancer Research.

Inflammation in atherosclerosis

Peter Libby

Cardiovascular Division, Department of Medicine, Brigham and Women's Hospital, and Harvard Medical School, Boston, Massachusetts 02115, USA (e-mail: plibby@rics.bwh.harvard.edu)

Abundant data link hypercholesterolaemia to atherogenesis. However, only recently have we appreciated that inflammatory mechanisms couple dyslipidaemia to atheroma formation. Leukocyte recruitment and expression of pro-inflammatory cytokines characterize early atherogenesis, and malfunction of inflammatory mediators mutes atheroma formation in mice. Moreover, inflammatory pathways promote thrombosis, a late and dreaded complication of atherosclerosis responsible for myocardial infarctions and most strokes. The new appreciation of the role of inflammation in atherosclerosis provides a mechanistic framework for understanding the clinical benefits of lipid-lowering therapies. Identifying the triggers for inflammation and unravelling the details of inflammatory pathways may eventually furnish new therapeutic targets.

Cardiovascular disease, currently the leading cause of death and illness in developed countries, will soon become the pre-eminent health problem worldwide¹. Atherosclerosis — a progressive disease characterized by the accumulation of lipids and fibrous elements in the large arteries — constitutes the single most important contributor to this growing burden of cardiovascular disease. Our views of the pathophysiology of this important malady have evolved substantively over the past century. The link between lipids and atherosclerosis dominated our thinking until the 1970s, based on strong experimental and clinical relationships between hypercholesterolaemia and atheroma². The emerging knowledge of vascular biology led to a focus on growth factors and the proliferation of smooth muscle cells in the 1970s and 1980s. The daunting clinical problem of restenosis (narrowing of the vessel lumen) following arterial intervention, considered a problem of proliferation, reinforced this interest in vascular growth control. A fusion of these views led to the concept of the atheroma as a graveyard of acellular lipid debris enrobed by a capsule of proliferated smooth muscle cells.

Over the past decade, however, we have come to appreciate a prominent role for inflammation in atherosclerosis and its complications. Whereas most clinicians previously regarded atheroma as a bland lesion, the current notion that inflammation and immune response contribute to atherogenesis has garnered increased interest³. As laboratory advances in vascular biology enabled new thinking about the clinical aspects of atherosclerosis, so too have emerging clinical data instructed our laboratory work, shifting its emphasis considerably. Formerly focused on luminal narrowing due to the bulk of atheroma, our current concepts recognize the biological attributes of the atheroma as key determinants of its clinical significance. This review will weave together laboratory and clinical advances to provide an update on inflammation in atherosclerosis.

Inflammation and the initiation of the atherosclerosis

The time-tested association of cholesterol with atherosclerosis stimulated a century-long study of the mechanisms linking lipids with atheroma. From the early years of the twentieth century onward, the pathogenesis of experimental atherosclerosis induced by hypercholesterolaemia has yielded to scrutiny at ever-deeper degrees of analysis. Indeed, instigation of inflammation may well link hyperlip-

idaemia to atherogenesis mechanistically. Soon after initiating an atherogenic diet, light microscopy reveals attachment of blood leukocytes to the endothelial cells that line the intima, the innermost layer of arteries⁴. Under ordinary circumstances, the endothelial monolayer in contact with flowing blood resists firm adhesion of leukocytes. We now possess considerable information about the molecular mechanisms of the attachment of white blood cells to endothelium. One endothelial-leukocyte adhesion molecule has emerged as a particularly attractive candidate for the early adhesion of mononuclear leukocytes to arterial endothelium at sites of atheroma initiation (Fig. 1). Vascular cell adhesion molecule-1 (VCAM-1) binds particularly those classes of leukocytes found in nascent atheroma: the monocyte and the T lymphocyte (Fig. 2).

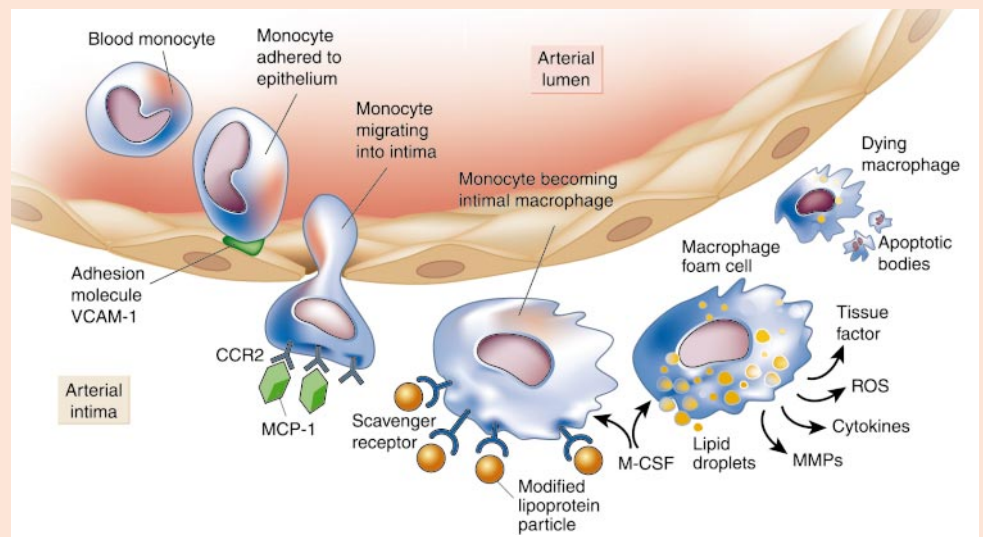
In addition to its leukocyte selectivity, other features of VCAM-1 make it an interesting candidate. Endothelial cells express VCAM-1 in response to cholesterol feeding selectively in areas prone to lesion formation⁵. In addition, VCAM-1 rises before leukocyte recruitment begins in both rabbit and mouse models of cholesterol-induced lesion formation⁶. Targeted deletion of VCAM-1 in mice causes embryonic lethality. However, experiments with hypomorphic variants of VCAM-1 introduced into mice rendered susceptible to atherogenesis (by inactivation of the apolipoprotein E (*apoE*) gene) show reduced lesion formation⁷. In addition to VCAM-1, P- and E-selectin also seem to contribute to leukocyte recruitment in atherosclerosis-susceptible mice^{8,9}.

The mechanism of VCAM-1 induction early after initiating an atherogenic diet probably depends on inflammation instigated by modified lipoprotein particles accumulating in the arterial intima in response to the hyperlipidaemia. Constituents of modified lipoprotein particles, among them certain oxidized phospholipids and short-chain aldehydes arising from lipoprotein oxidation, can induce transcriptional activation of the VCAM-1 gene mediated in part by nuclear factor- κ B (NF- κ B)¹⁰. Pro-inflammatory cytokines such as interleukin (IL)-1 β or tumour-necrosis factor- α (TNF- α) induce VCAM-1 expression in endothelial cells by this pathway. Human atherosclerotic lesions contain these cytokines. Thus, pro-inflammatory cytokines may link hypercholesterolaemia to VCAM-1 expression.

Endogenous anti-inflammatory pathways and 'atheroprotection'

The mechanism of focal expression of VCAM-1 selectively in sites of lesion formation has been the subject of intense recent investigation. One novel idea to emerge from experi-

Figure 1 Mononuclear phagocytes in atherogenesis. This figure schematizes steps in the recruitment of mononuclear phagocytes to the nascent atherosclerotic plaque and some of the functions of these cells in the mature atheroma. The steps are depicted in an approximate time sequence proceeding from left to right. The normal arterial endothelium resists prolonged contact with leukocytes including the blood monocyte. When endothelial cells undergo inflammatory activation, they increase their expression of various leukocyte adhesion molecules. In the context of monocyte recruitment to the atheroma, vascular cell adhesion molecule-1 (VCAM-1) seems to have a major role. Once adherent to the activated



endothelial layer, the monocyte diapedes between intact endothelial cells to penetrate into the tunica intima, or innermost layer of the arterial wall. This directed migration requires a chemoattractant gradient. Various chemokines seem to participate in this process, particularly interaction of monocyte chemoattractant protein-1 (MCP-1) with its receptor CCR2. Once resident in the intima the monocyte acquires characteristics of the tissue macrophage. In the atheroma in particular, the macrophage expresses scavenger receptors that bind internalized lipoprotein particles modified for example by oxidation or glycation. These processes give rise to the arterial foam cell, a hallmark of the arterial lesion, so named because of the foamy appearance under the microscope, which is the result of accumulation of lipid droplets within the cytoplasm. Within the arterial intima, the macrophage serves many functions related to atherosclerosis and its complications. Notably, the foam cell secretes pro-inflammatory cytokines that amplify the local inflammatory response in the lesion, as well as reactive oxygen species. The activated mononuclear phagocyte has a key role in the thrombotic complications of atherosclerosis by producing matrix metalloproteinases (MMPs) that can degrade extracellular matrix that lends strength to the plaque's fibrous cap. When the plaque ruptures as a consequence, it permits the blood to contact another macrophage product, the potent pro-coagulant protein tissue factor. Eventually the macrophages congregate in a central core in the typical atherosclerotic plaque. Macrophages can die in this location, some by apoptosis, hence producing the so-called 'necrotic core' of the atherosclerotic lesion.

mental work — 'atheroprotection' — stands the traditional view on its head. Rather than asking what goes awry at sites of lesion formation, one can reverse the question and ask what qualities of endothelium in unaffected areas confer resistance to lesion initiation. Regions of the arterial tree protected from atherosclerosis usually experience laminar shear stress due to orderly blood flow. Sites predisposed to lesion formation include branch points of arteries, which experience disturbed rather than laminar flow.

A number of genes with potentially 'atheroprotective' properties contain shear-stress response elements in their promoter regions. Many such atheroprotective genes may modulate inflammation. For example, superoxide dismutase, expressed at higher levels in regions of laminar flow, may combat oxidative stress and hence limit VCAM-1 expression and other inflammatory pathways¹¹. Likewise, nitric oxide arising from endothelial nitric oxide synthase, another shear stress-regulated gene, can inhibit VCAM gene expression through a novel pathway involving inhibition of the activation of NF- κ B, the central transcriptional control point in vascular inflammation¹². These new insights from the laboratory provide a potential explanation for the tendency of atheroma to form in characteristic sites of flow disturbance in the arterial tree despite similar exposure to fluid-phase risk factors such as hypercholesterolaemia.

Mechanisms of leukocyte chemoattraction

Morphologic studies have established that, once adherent to the endothelial cell, leukocytes enter the intima by diapedesis between endothelial cells at their junctions. This phenomenon of directed migration of leukocytes through the endothelium, known for well over a century, has in the past few years yielded to molecular analysis. Investigators have defined families of chemoattractant cytokines (chemokines) capable of recruiting leukocytes into the arterial intima. For example, monocyte chemoattractant protein-1 (MCP-1), overexpressed in human and experimental atheroma, can recruit the mononuclear phagocytes that characteristically accumulate in the

nascent atheroma (Fig. 1). Recent work using compound mutant mice lacking MCP-1 or its receptor CCR2, and susceptible to atherosclerosis owing to the absence of genes encoding apoE or the low-density lipoprotein (LDL) receptor, has shown striking decreases in mononuclear phagocyte accumulation and local lipid levels^{13,14}. IL-8 may have a similar role as a leukocyte chemoattractant during atherogenesis¹⁵. Atheroma overexpress other chemokines that may contribute to lymphocyte recruitment, including a trio of CXC chemokines induced by interferon- γ (IFN- γ)¹⁶ (Fig. 2). Chemoattraction of mast cells found in atheroma may depend on eotaxin, a CC chemokine also overexpressed in these lesions¹⁷ (Fig. 3).

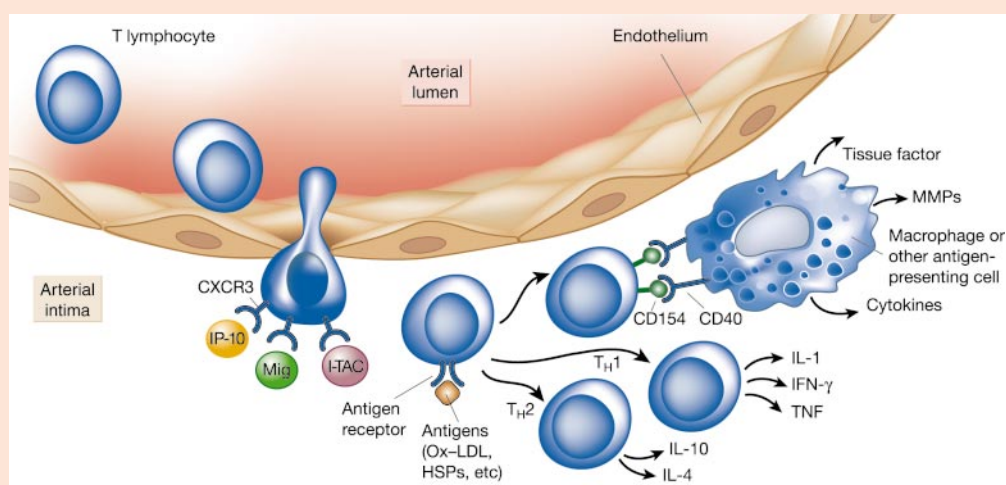
Mechanisms of leukocyte activation in the intima

Once resident in the arterial intima, monocytes acquire the morphological characteristics of macrophages, undergoing a series of changes that lead ultimately to foam cell formation. The monocytes increase expression of scavenger receptors from modified lipoproteins such as the scavenger receptor A (SRA) and CD36, and then internalize modified lipoproteins, such that cholesterol esters accumulate in cytoplasmic droplets (Fig. 1). These lipid-laden macrophages, known as foam cells, characterize the early atherosclerotic lesion. Macrophages within atheroma also secrete a number of growth factors and cytokines involved in lesion progression and complication (see below). In addition, macrophages replicate within the intima.

Studies performed a decade ago identified macrophage colony-stimulating factor (M-CSF) as a candidate activator of several of the steps that stimulate transition of the monocyte to the lipid-laden macrophage. M-CSF augments SRA expression, increases production of cytokines and growth factors by these cells, and also serves as a survival and co-mitogenic stimulus. Both experimental and human atherosclerotic plaques overexpress M-CSF^{18,19}.

Studies of mice with mutations that inactivate M-CSF, bred onto atherosclerosis-susceptible backgrounds, permitted direct testing of

Figure 2 The roles of T lymphocytes in atherogenesis. As in the case of the mononuclear phagocyte, lymphocytes enter the intima facilitated by binding to adhesion molecules including vascular cell adhesion molecule-1 (VCAM-1) and in response to chemoattractants selective for lymphocytes. Known chemoattractants include a trio of interferon- γ (IFN- γ)-inducible chemokines of the CXC family including inducible protein-10 (IP-10), monokine induced by IFN- γ (Mig), and IFN-inducible T-cell α -chemoattractant (I-TAC). These chemokines bind to chemokine receptor CXCR3 expressed by T cells in the atherosclerotic lesion. Once resident in the arterial intima, the T cell may encounter antigens such as oxidized low-density lipoprotein (Ox-LDL) and heat-shock proteins (HSPs) of endogenous or microbial origin, among others. Upon activation by engagement of the receptor and antigen, the T cell can produce cytokines that can influence the behaviour of other cells present in the atheroma. Notably, CD154 binding to CD40 ligand, particularly on macrophages, may induce the expression of tissue factor, matrix metalloproteinases (MMPs) and pro-inflammatory cytokines. The production of these mediators provides an amplification loop resulting from crosstalk between the prototypical cell of acquired immunity (the T lymphocyte) and that of innate immunity (the mononuclear phagocyte). Within the atheroma, as in other tissues, the helper T cells can polarize into those secreting generally pro-inflammatory cytokines (known as T_H1 cells) and/or those secreting predominantly anti-inflammatory cytokines (denoted T_H2 cells). In general, T_H1 cells predominate in the atheroma. But experimental data in mice suggest that with extreme levels of hypercholesterolaemia the balance may shift towards T_H2 predominance. Recent evidence indicates that in abdominal aortic aneurysms, T_H2 cytokines predominate in contrast with the situation in occlusive atherosclerotic disease.



the role of M-CSF in the formation of atheromatous lesions. Mice lacking M-CSF show retarded lesion development with markedly reduced macrophage accumulation^{20,21}. This effect occurred in mice lacking both apoE and the LDL receptor and depended on gene dosage²². Granulocyte-macrophage colony-stimulating factor (GM-CSF) may also promote inflammation in the atheroma. GM-CSF aids the survival of a population of mononuclear phagocytes that contain myeloperoxidase, an enzyme that gives rise to the pro-oxidant hypochlorous acid, a potential source of oxidative stress and inflammation in the human plaque²³.

These examples illustrate how specific candidates identified by descriptive studies have proven causally related to inflammation during atherogenesis using genetically altered mice. From the adherence to VCAM-1, to the chemoattractant response to MCP-1, to the activation by M-CSF, we are now beginning to understand the mechanisms by which mononuclear phagocytes and inflammatory signalling pathways participate in formation of the fatty streak, the initial lesion of atherosclerosis (Fig. 1).

Inflammation in atheroma progression and complication

After formation of the fatty streak, the nascent atheroma typically evolves into a more complex lesion, which eventually leads to clinical manifestations. Although past discussions neatly separated the progression and complication phases of atherosclerosis, we now recognize the blurred barriers between these different aspects of atherogenesis.

According to the traditional notion, fatty streaks evolve into complicated atheroma through multiplication of smooth muscle cells, which accumulate in the plaque and lay down an abundant extracellular matrix. As the lesion becomes more bulky, the arterial lumen narrows until it hampers flow and leads to clinical manifestations: in the coronary circulation, unstable *angina pectoris*, or acute myocardial infarction. Growth factors elaborated by macrophages in the atherosclerotic intima supposedly stimulated the smooth muscle replication responsible for lesion growth. According to the classical view, this process occurred in an inevitable and progressive fashion gradually during time.

Plaque disruption and discontinuous progression of atheroma

Clinical observations have challenged the concept of continuous growth of atheroma, prompting a re-evaluation of the biology thought to underlie atheroma progression. Data that emerged from serial angiographic studies suggest that many coronary arterial lesions in humans develop stenoses discontinuously. In patient populations successively undergoing angiography at three different times, smooth progression of the lesions proved the exception rather than the rule^{24,25}.

What might explain the apparent 'bursts' in growth of atheroma in these studies in humans? Observations on the microscopic pathology of atherosclerotic plaques provided clues. Current evidence suggests that physical disruption of plaques may trigger thrombosis and thus promote sudden expansion of atheromatous lesions²⁶. Three types of physical disruption may occur²⁷.

Superficial erosion, or microscopic areas of desquamation of endothelial cells that form the monolayer covering the intima, occurs frequently in both humans and animals with experimentally induced atherosclerosis. Such areas of limited endothelial desquamation often form the nidus of a platelet thrombus as they uncover sub-endothelial collagen and von Willebrand factor that promote platelet adhesion and activation²⁸. Although common and most often asymptomatic, such superficial erosion may account for approximately one-quarter of fatal coronary thromboses.

Disruption of the microvessels that form in atherosclerotic plaques furnishes another scenario for sudden plaque progression²⁹. Atheromata develop microvascular channels as a result of neo-angiogenesis. Like those that form in the diabetic retina, the new blood vessels in the plaque may be particularly fragile and prone to micro-haemorrhage. Multiple lines of evidence support thrombosis *in situ* within plaques during human atherogenesis. Intra-plaque deposition of fibrin and fibrin-split products and haemosiderin provide evidence of intra-plaque haemorrhage. The thrombosis *in situ* leads to thrombin generation, which, in addition to cleaving fibrinogen, can potently stimulate smooth muscle migration and proliferation. Thrombin triggers platelet release of growth factors such as platelet-derived growth factor (PDGF) from their alpha

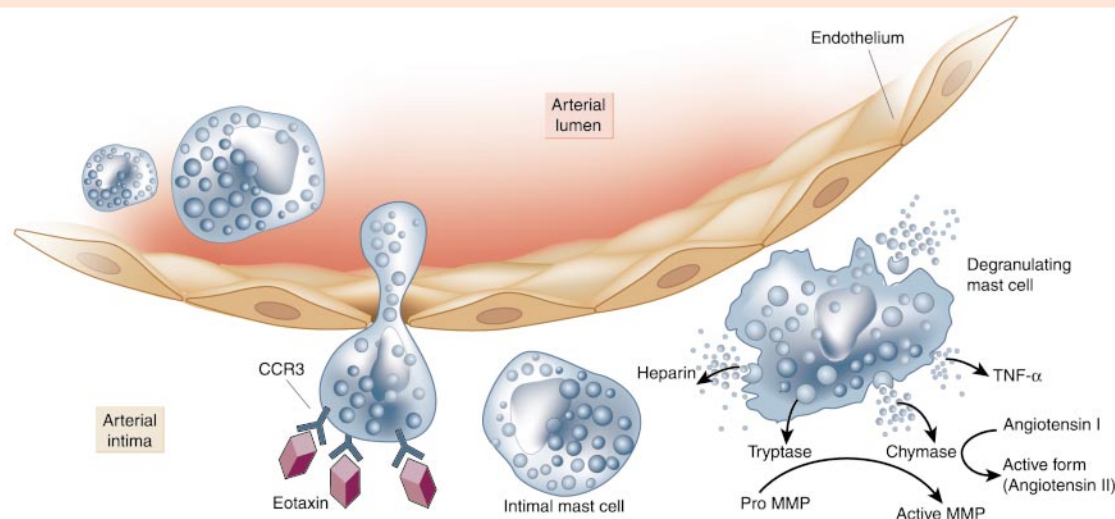


Figure 3 Recruitment and functions of mast cells in atherogenesis. The leukocytic infiltrate within atheromatous plaques includes a small but potentially important population of mast cells. Eotaxin, a chemoattractant that interacts with the chemokine receptor CCR3, may mediate the trans-endothelial migration of this specialized leukocyte. Once resident in the intima, the mast cell can undergo degranulation, releasing preformed tumour-necrosis factor- α (TNF- α), heparin with its anti-coagulant and potentially growth inhibitory effects on smooth muscle cells, and the serine proteinases tryptase and chymase. These proteinases may activate the inactive zymogen forms of matrix metalloproteinases (MMPs) to their proteolytic forms. Chymase may also generate active forms of angiotensin from their precursor, angiotensin I.

granules, further stimulating smooth muscle migration and proliferation. Activated platelets also elaborate transforming growth factor beta (TGF- β), the most potent stimulus known for interstitial collagen synthesis by smooth muscle cells. In this manner, a silent microvascular haemorrhage within the atherosclerotic intima could give rise to a growth spurt in the evolution of the plaque.

The third and most common mechanism of plaque disruption, a fracture of the plaque's fibrous cap, also involves inflammation (Fig. 4). The plaque's fibrous cap usually serves to sequester the thrombogenic lipid-rich core of the atheroma from the bloodstream, which contains circulating coagulation proteins. Fissure of the fibrous cap allows the coagulation factors contact with tissue factor, the main pro-thrombotic stimulus found in the lesion's lipid core. Although the ruptured fibrous cap causes some three-quarters of acute myocardial infarctions, like the other forms of plaque disruption, most episodes probably cause no clinical symptoms. When the prevailing fibrinolytic mechanisms outweigh the pro-coagulant pathways, a limited mural thrombus, rather than an occlusive and sustained blood clot, forms. With healing, however, resorption of the mural thrombus and the release of PDGF and the anti-inflammatory mediator TGF- β combine to engender a healing response that leads to fibrous tissue formation. The consequent smooth muscle accumulation and collagen accretion allow rapid evolution of a fatty lesion to one of more fibrous character (Fig. 4).

These examples illustrate the inextricable links between thrombosis and lesion progression. Usually below the clinical threshold, evolution of the lesion most often occurs silently, leading to transition from the fatty to the fibrous atherosclerotic plaque.

Inflammation causes various forms of plaque disruption

We know little of the mechanisms of superficial erosion of atherosclerotic plaques. Two processes related to inflammation may participate in endothelial desquamation. The first, endothelial cell death (perhaps by apoptosis) may result from local production of inflammatory mediators or cytolytic attack by activated killer T cells. Additionally, inflammatory mediators and oxidized lipoproteins can stimulate the expression and activation of matrix metalloproteinases (MMPs) specialized in degrading components of the sub-endothelial basement membrane³⁰. Thus, inflammatory stimulation may

promote the production by endothelial cells of enzymes that degrade the extracellular matrix constituents to which they adhere under normal circumstances. In this fashion, inflammation can promote loss of endothelium, the hallmark of superficial erosion.

The mechanisms of microvessel formation in atheroma probably resemble those common to other sites of angiogenesis. In addition to secreting growth factors for smooth muscle cells, inflammatory cells residing in the plaque, including macrophages, produce angiogenic mediators such as acidic and basic fibroblast growth factor and vascular endothelial growth factor (VEGF)^{31,32}. Microvessels in plaques may not only serve as a site for haemorrhage *in situ* and thrombosis, but may also perform a nutritive function promoting plaque growth. Indeed, administration of inhibitors of angiogenesis retards microvessel formation and lesion evolution in atherosclerosis-prone mice³³. The plaque microvasculature may therefore promote lesion evolution in two ways. The potential adverse effects of promoting plaque angiogenesis require consideration when contemplating strategies for promoting therapeutic angiogenesis in ischaemic hearts.

Among the forms of plaque disruption, we best understand fracture of the fibrous cap³⁴. Interstitial collagen molecules confer most of the tensile strength on the fibrous cap³⁵, and several tightly regulated processes determine the level of collagen crucial for stability of this structure. Certain pro-inflammatory cytokines, such as IFN- γ , can inhibit collagen production by smooth muscle cells, the principle source of this extracellular matrix macromolecule in the arterial wall. Interstitial collagen fibrils usually resist proteolytic degradation, and only a limited number of interstitial collagenases can make an initial proteolytic nick in the collagen chains that make up the triple helical collagen fibril. We have found overexpression of all three human interstitial collagenases in atheromatous plaques (MMP-1, -8 and -13)³⁶⁻³⁸.

After the limited proteolytic cleavage arising from the action of interstitial collagenases, gelatinases continue collagen catabolism. Extracts of atheroma show augmented active forms of two gelatinases (MMP-2 and MMP-9)³⁶. Arteries do express the endogenous antagonists of MMPs, the tissue inhibitors of metalloproteinases (TIMPs). However, evidence for collagenolysis *in situ* indicates excess active forms of interstitial collagenases over the TIMPs in

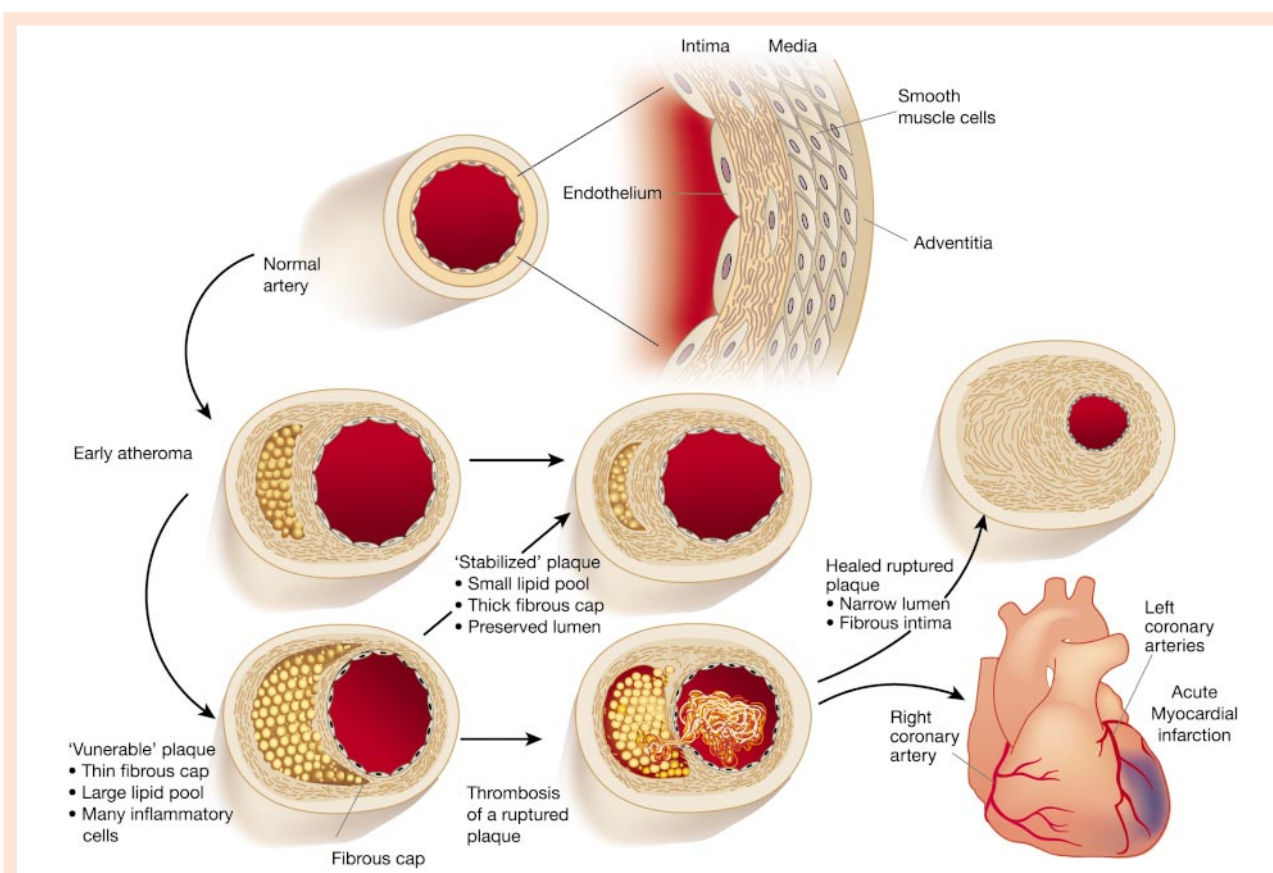


Figure 4 Schematic of the life history of an atheroma. The normal human coronary artery has a typical trilaminar structure. The endothelial cells in contact with the blood in the arterial lumen rest upon a basement membrane. The intimal layer in adult humans generally contains a smattering of smooth muscle cells scattered within the intimal extracellular matrix. The internal elastic lamina forms the barrier between the tunica intima and the underlying tunica media. The media consists of multiple layers of smooth muscle cells, much more tightly packed than in the diffusely thickened intima, and embedded in a matrix rich in elastin as well as collagen. In early atherogenesis, recruitment of inflammatory cells (Figs 1–3) and the accumulation of lipids leads to formation of a lipid-rich core, as the artery enlarges in an outward, abluminal direction to accommodate the expansion of the intima. If inflammatory conditions prevail and risk factors such as dyslipidaemia persist, the lipid core can grow, and proteinases secreted by the activated leukocytes can degrade the extracellular matrix, while pro-inflammatory cytokines such as interferon- γ (IFN- γ) can limit the synthesis of new collagen. These changes can thin the fibrous cap and render it friable and susceptible to rupture. When the plaque ruptures, blood comes in contact with the tissue factor in the plaque coagulates. Platelets activated by thrombin generated from the coagulation cascade and by contact with the intimal compartment instigate thrombus formation. If the thrombus occludes the vessel persistently, an acute myocardial infarction can result (the dusky blue area in the anterior wall of the left ventricle, lower right). The thrombus may eventually resorb as a result of endogenous or therapeutic thrombolysis. However, a wound healing response triggered by thrombin generated during blood coagulation can stimulate smooth muscle proliferation. Platelet-derived growth factor (PDGF) released from activated platelets stimulates smooth muscle cell migration. Transforming growth factor- β (TGF- β), also released from activated platelets, stimulates interstitial collagen production. This increased migration, proliferation and extracellular matrix synthesis by smooth muscle cells thickens the fibrous cap and causes further expansion of the intima, often now in an inward direction, yielding constriction of the lumen. Stenotic lesions produced by the luminal encroachment of the fibrosed plaque may restrict flow, particularly under situations of increased cardiac demand, leading to ischaemia, commonly provoking symptoms such as *angina pectoris*. Advanced stenotic plaques, being more fibrous, may prove less susceptible to rupture and renewed thrombosis. Lipid lowering can reduce lipid content and calm the intimal inflammatory response, yielding a more 'stable' plaque with a thick fibrous cap and a preserved lumen (centre).

human atherosclerotic plaques³⁷. *In vitro* studies have shown that inflammatory mediators found in atheroma, such as IL-1 β , TNF- α , and CD40 ligand (CD154), augment MMP expression in mononuclear phagocytes and endothelial and smooth muscle cells. Mast cells in the lesion may release the MMP inducer TNF- α as well as serine proteinases that can activate latent MMP proenzymes^{39,40} (Fig. 3).

Converging lines of evidence point to the dynamic regulation of collagen levels in the plaque's fibrous cap. When inflammation prevails in the intima, smooth muscle cell production of new collagen required for repair and maintenance of the fibrous cap decreases. Meanwhile, collagen degradation increases due to overexpression of active MMPs. The net result, dissolution of the collagenous matrix of the fibrous cap, renders this structure weak, friable and susceptible to fracture when exposed to haemodynamic stresses. Indeed, pathologists categorize plaques as those exhibiting signs of stability, notably a

thick fibrous cap, and those prone to rupture, having a thin fibrous cap and a scant collagenous skeleton on pathological examination.

Triggers for inflammation

Although the concept that inflammation occupies a central position in the pathophysiology of atherosclerosis has gained considerable currency, knowledge of the inciting factors remains remarkably sketchy. Much of the progress in understanding atherosclerosis over the past 50 years has depended on the lipid hypothesis. LDL cholesterol undoubtedly contributes importantly to atherosclerosis in many cases, and may indeed constitute a ubiquitous permissive factor for atherogenesis. However, most individuals with proven coronary artery disease in the United States have 'average' levels of cholesterol. ('Average' levels of cholesterol in developed countries probably exceed by far truly normal levels for our species as suggested by extrapolation

from data on animals and humans in agrarian societies.) Even extremely effective therapies targeting LDL cholesterol reduce coronary events by at most one-third over a five-year treatment period. Earlier or longer lipid-lowering therapy might further reduce the residual risk of atherosclerotic disease. However, addressing risk factors other than LDL cholesterol may also ameliorate atherosclerosis.

The strength of evidence supporting 'non-traditional', emerging risk factors in atherogenesis currently lags behind cholesterol, and further study is required to clarify their role. Examples of novel risk factors include lipoprotein (a), homocysteine, infectious agents such as herpesvirus and *Chlamydia pneumoniae*, and oxidant stress evoked by the pressor hormone angiotensin II. The view of angiotensin II as a pro-inflammatory and pro-oxidant stimulus furnishes a satisfying link between the mechanisms at play in hypertension and its common companion, atherosclerosis.

As an epidemic of obesity sweeps the world, with insulin resistance and diabetes close behind, the so-called 'metabolic syndrome' has emerged as one of the main contributors to risk for atherosclerosis. Adipose tissue itself can give rise to cytokines that worsen insulin sensitivity, and provide a systemic pro-inflammatory stimulus. In the metabolic syndrome, LDL levels often remain in the average range, although the particles may have qualitative alterations that render them small and dense, making them particularly prone to oxidation and hence evoking inflammation. The low levels of high-density lipoprotein (HDL) that characteristically accompany the elevated triglycerides in the metabolic syndrome blunt another endogenous anti-inflammatory and hence atheroprotective mechanism⁴¹. HDL particles may owe their protection against atherosclerosis not only to reverse cholesterol transport, but also to provision of antioxidant enzymes such as paraoxonase and platelet-activating factor acetyl hydrolase. Persistent hyperglycaemia in diabetes can accelerate the formation of advanced glycation end-products, yet another trigger to arterial inflammation⁴². Thus, in addition to LDL, many putative non-traditional factors may aggravate atherogenesis by promoting inflammation.

Inflammation as a therapeutic target in atherosclerosis

Our new understanding of the pivotal position of inflammation in the pathogenesis of atherosclerosis raises questions and opens opportunities in prevention and therapy of this disease. A series of large, well-designed, randomized and controlled clinical trials have recently established the utility of several different pharmacological strategies for preventing recurrent myocardial infarction or death beyond the recognized roles of aspirin and β -adrenergic blocking agents. Newer drug classes shown to be effective in this regard, and listed in decreasing order of the strength of evidence, include inhibitors of hydroxymethylglutaryl coenzyme A (statins); angiotensin-converting enzyme inhibitors and angiotensin-receptor blockers; and fibric acid derivatives (activators of the nuclear receptor/transcription factor peroxisome proliferator-activated receptor- α or PPAR- α). The success of these categories of agents in the clinic has prompted intense investigation, in the context of inflammation biology in atherosclerosis, to garner a more complete picture of the mechanism(s) of the clinical benefit observed.

For example, statins not only inhibit cholesterol synthesis, but also block the production of isoprenoid intermediates such as farnesyl- or geranylgeranyl-pyrophosphates, which are important in modifying small G proteins, among other biochemical effects. A number of laboratory studies have addressed the hypothesis that the non-lipid-lowering effects of statins may contribute to their clinical benefit. The possible 'pleiotropic' effects of this class of drugs include anti-inflammatory actions such as reduction in leukocyte adhesion, and antagonizing aspects of macrophage activation including replication, metalloproteinase production, and tissue factor procoagulant gene expression⁴³.

The degree to which certain clinical benefits of statins derive from such direct anti-inflammatory effects remains controversial. Many of

the *in vitro* studies that demonstrate statin-induced reduction in pro-atherogenic functions of isolated cells have used concentrations of these agents not likely to be achieved in tissues clinically. In addition, pravastatin, which is relatively cell-impermeant owing to its hydrophilicity compared to most other statins, lacks such *in vitro* effects, but has proven effective in reducing cardiovascular events in multiple clinical trials. Statins certainly do stem inflammation in patients with atheroma, as gauged by the marker C-reactive protein (CRP)⁴⁴. However, the degree of lowering of CRP correlates poorly with a patient's drop in LDL, hinting that some of the anti-inflammatory effect may not derive simply from a lipid-lowering action.

Just as reduced LDL may not account for all of the benefits of statins, recent clinical trials suggest benefits of interrupting angiotensin II signalling that are not accounted for by the degree of blood pressure lowering^{45,46}. Indeed, angiotensin II's actions extend far beyond vasoconstriction. Considerable evidence now supports a role for angiotensin II as a pro-inflammatory mediator, elevating it to the category of an 'honorary' cytokine⁴⁷. For example, this peptide can elicit VCAM-1 and MCP-1 expression by endothelial cells, and IL-6 production by smooth muscle cells.

The recent clinical success of fibric acid derivatives in certain patient populations, including those with diabetes or diabetic-like insulin-resistant states, has stimulated intense interest in the PPAR- α pathway. PPAR- α agonism increases the synthesis of apoA1, the main apoprotein of HDL, a particle that protects against lesion formation, probably owing to its role in reverse cholesterol transport (removing cholesterol from the artery wall and delivering it to peripheral tissues and the liver). Other laboratory studies have established that PPAR- α agonists also possess anti-inflammatory properties of potential relevance to atherogenesis. For example, these agents can reduce VCAM-1 and tissue factor gene expression by cells found in atheroma^{48–50}. Interference with the activation of NF- κ B, resulting from competition for co-activators, may explain part of this anti-inflammatory action of PPAR- α agonism⁵¹.

These examples provide illustrations of unexpected anti-inflammatory effects of existing therapies for atherosclerosis. Uncovering inflammatory pathways has raised the possibility that future treatments may target effectors of inflammation directly to add to the benefit of current treatments. Potential targets include proximal triggers such as infectious agents, central signalling hubs in inflammation such as NF- κ B, and distal effectors such as MMPs, adhesion molecules, and the like. Targeting NF- κ B transcription pathways for a chronic disease such as atherosclerosis may well prove impractical given the key role of inflammation and innate immunity in normal host defences. The redundancy of distal effectors of inflammation suggests to me that narrow-spectrum inhibition may not effectively modify the disease process, while broad blockade of these mediators will impair host defences much as would interruption of NF- κ B activation. I foresee targeting the proximal triggers as the most promising strategy for interrupting inflammation in atherogenesis.

Inflammatory markers as gauges of atherosclerotic risk

As noted above, many individuals develop coronary heart disease in the absence of abnormalities in the lipoprotein profile. The availability of effective therapies for preventing even a first myocardial infarction renders imperative the need to identify individuals at risk for concerted intervention before problems manifest. Based on the evidence supporting a role for inflammation in the pathogenesis of atherosclerosis, serum markers of inflammation have garnered substantial interest as markers of atherosclerotic risk; these add to the information available from traditional measures such as the lipid profile.

One of these markers, CRP, has proven remarkably robust as a marker of cardiovascular risk. Plasma CRP, an acute phase reactant produced primarily by the liver in response to inflammatory cytokines such as IL-6, prospectively identifies asymptomatic individuals at risk for coronary events. Although many candidates as novel markers of risk exist, they must meet several criteria to prove

clinically useful. The marker must have a rigorously standardized and reproducible assay, be relatively stable from day to day in a given individual, and add to estimates of risk provided by established markers such as the lipid profile as determined in prospective study. The promise of CRP in this regard has engendered clinical trials that will test its ability to guide preventive therapy in apparently well individuals. We therefore stand on the threshold of clinical application of the basic biology of inflammation in atherosclerosis that could fundamentally alter the way in which we practice preventive medicine and prove immeasurably beneficial to the public as well. □

doi:10.1038/nature01323

1. Murray, C. J. & Lopez, A. D. Global mortality, disability, and the contribution of risk factors: Global Burden of Disease Study. *Lancet* **349**, 1436–1442 (1997).
2. Ross, R. & Harker, L. Hyperlipidemia and atherosclerosis. *Science* **193**, 1094–1100 (1976).
3. Libby, P., Ridker, P. M. & Maseri, A. Inflammation and atherosclerosis. *Circulation* **105**, 1135–1143 (2002).
4. Poole, J. C. F. & Florey, H. W. Changes in the endothelium of the aorta and the behavior of macrophages in experimental atheroma of rabbits. *J. Pathol. Bacteriol.* **75**, 245–253 (1958).
5. Cybulsky, M. I. & Gimbrone, M. A. Jr Endothelial expression of a mononuclear leukocyte adhesion molecule during atherogenesis. *Science* **251**, 788–791 (1991).
6. Li, H., Cybulsky, M. I., Gimbrone, M. A. Jr & Libby, P. An atherogenic diet rapidly induces VCAM-1, a cytokine-regulatable mononuclear leukocyte adhesion molecule, in rabbit endothelium. *Arterioscler. Thromb.* **13**, 197–204 (1993).
7. Cybulsky, M. I. *et al.* A major role for VCAM-1, but not ICAM-1, in early atherosclerosis. *J. Clin. Invest.* **107**, 1255–1262 (2001).
8. Johnson, R. C. *et al.* Absence of P-selectin delays fatty streak formation in mice. *J. Clin. Invest.* **99**, 1037–1043 (1997).
9. Dong, Z. M. *et al.* The combined role of P- and E-selectins in atherosclerosis. *J. Clin. Invest.* **102**, 145–152 (1998).
10. Collins, T. & Cybulsky, M. I. NF- κ B: pivotal mediator or innocent bystander in atherogenesis? *J. Clin. Invest.* **107**, 255–264 (2001).
11. Topper, J. N. & Gimbrone, M. A. Jr Blood flow and vascular gene expression: fluid shear stress as a modulator of endothelial phenotype. *Mol. Med. Today* **5**, 40–46 (1999).
12. De Caterina, R. *et al.* Nitric oxide decreases cytokine-induced endothelial activation. Nitric oxide selectively reduces endothelial expression of adhesion molecules and proinflammatory cytokines. *J. Clin. Invest.* **96**, 60–68 (1995).
13. Gu, L. *et al.* Absence of monocyte chemoattractant protein-1 reduces atherosclerosis in low-density lipoprotein-deficient mice. *Mol. Cell* **2**, 275–281 (1998).
14. Boring, L., Gosling, J., Cleary, M. & Charo, I. F. Decreased lesion formation in CCR2^{-/-} mice reveals a role for chemokines in the initiation of atherosclerosis. *Nature* **394**, 894–897 (1998).
15. Boisvert, W. A., Santiago, R., Curtiss, L. K. & Terkeltaub, R. A. A leukocyte homologue of the IL-8 receptor CXCR-2 mediates the accumulation of macrophages in atherosclerotic lesions of LDL receptor-deficient mice. *J. Clin. Invest.* **101**, 353–363 (1998).
16. Mach, F. *et al.* Differential expression of three T lymphocyte-activating CXC chemokines by human atheroma-associated cells. *J. Clin. Invest.* **104**, 1041–1050 (1999).
17. Haley, K. J. *et al.* Overexpression of eotaxin and the CCR3 receptor in human atherosclerosis: using genomic technology to identify a potential novel pathway of vascular. *Circulation* **102**, 2185–2189 (2000).
18. Clinton, S., Underwood, R., Sherman, M., Kufe, D. & Libby, P. Macrophage-colony stimulating factor gene expression in vascular cells and in experimental and human atherosclerosis. *Am. J. Pathol.* **140**, 301–316 (1992).
19. Rosenfeld, M. *et al.* Macrophage colony-stimulating factor mRNA and protein in atherosclerotic lesions of rabbits and humans. *Am. J. Pathol.* **140**, 291–300 (1992).
20. Smith, J. D. *et al.* Decreased atherosclerosis in mice deficient in both macrophage colony-stimulating factor (op) and apolipoprotein E. *Proc. Natl Acad. Sci. USA* **92**, 8264–8268 (1995).
21. Rajavashisth, T. *et al.* Heterozygous osteopetrotic (op) mutation reduces atherosclerosis in LDL receptor-deficient mice. *J. Clin. Invest.* **101**, 2702–2710 (1998).
22. Qiao, J. H. *et al.* Role of macrophage colony-stimulating factor in atherosclerosis: studies of osteopetrotic mice. *Am. J. Pathol.* **150**, 1687–1699 (1997).
23. Sugiyama, S. *et al.* Macrophage myeloperoxidase regulation by granulocyte macrophage colony-stimulating factor in human atherosclerosis and implications in acute coronary syndromes. *Am. J. Pathol.* **158**, 879–891 (2001).
24. Bruschke, A. V. *et al.* The dynamics of progression of coronary atherosclerosis studied in 168 medically treated patients who underwent coronary arteriography three times. *Am. Heart J.* **117**, 296–305 (1989).
25. Yokoyama, K. *et al.* Process of progression of coronary artery lesions from mild or moderate stenosis to moderate or severe stenosis: a study based on four serial coronary arteriograms per year. *Circulation* **100**, 903–909 (1999).
26. Davies, M. J. Stability and instability: the two faces of coronary atherosclerosis. The Paul Dudley White Lecture, 1995. *Circulation* **94**, 2013–2020 (1996).
27. Virmani, R., Burke, A. P., Farb, A. & Kolodgie, F. D. Pathology of the unstable plaque. *Prog. Cardiovasc. Dis.* **44**, 349–356 (2002).
28. Faggiotto, A., Ross, R. & Harker, L. Studies of hypercholesterolemia in the nonhuman primate. I. Changes that lead to fatty streak formation. *Arteriosclerosis* **4**, 323–340 (1984).
29. de Boer, O. J., van der Wal, A. C., Teeling, P. & Becker, A. E. Leucocyte recruitment in rupture prone regions of lipid-rich plaques: a prominent role for neovascularization? *Cardiovasc. Res.* **41**, 443–449 (1999).
30. Rajavashisth, T. B. *et al.* Inflammatory cytokines and oxidized low density lipoproteins increase endothelial cell expression of membrane type 1-matrix metalloproteinase. *J. Biol. Chem.* **274**, 11924–11929 (1999).
31. Brogi, E. *et al.* Distinct patterns of expression of fibroblast growth factors and their receptors in human atheroma and non-atherosclerotic arteries: association of acidic FGF with plaque microvessels and macrophages. *J. Clin. Invest.* **92**, 2408–2418 (1993).
32. Ramos, M. A. *et al.* Induction of macrophage VEGF in response to oxidized LDL and VEGF accumulation in human atherosclerotic lesions. *Arterioscler. Thromb. Vasc. Biol.* **18**, 1188–1196 (1998).
33. Moulton, K. S. *et al.* Angiogenesis inhibitors endostatin or TNP-470 reduce intimal neovascularization and plaque growth in apolipoprotein E-deficient mice. *Circulation* **99**, 1726–1732 (1999).
34. Libby, P. The molecular bases of the acute coronary syndromes. *Circulation* **91**, 2844–2850 (1995).
35. Lee, R. & Libby, P. The unstable atheroma. *Arterioscler. Thromb. Vasc. Biol.* **17**, 1859–1867 (1997).
36. Galis, Z., Sukhova, G., Lark, M. & Libby, P. Increased expression of matrix metalloproteinases and matrix degrading activity in vulnerable regions of human atherosclerotic plaques. *J. Clin. Invest.* **94**, 2493–2503 (1994).
37. Sukhova, G. K. *et al.* Evidence for increased collagenolysis by interstitial collagenases-1 and -3 in vulnerable human atheromatous plaques. *Circulation* **99**, 2503–2509 (1999).
38. Herman, M. P. *et al.* Expression of neutrophil collagenase (matrix metalloproteinase-8) in human atheroma: a novel collagenolytic pathway suggested by transcriptional profiling. *Circulation* **104**, 1899–1904 (2001).
39. Saren, P., Welgus, H. G. & Kovanen, P. T. TNF- α and IL-1 β selectively induce expression of 92-kDa gelatinase by human macrophages. *J. Immunol.* **157**, 4159–4165 (1996).
40. Kovanen, P. T., Kaartinen, M. & Paavonen, T. Infiltrates of activated mast cells at the site of coronary atheromatous erosion or rupture in myocardial infarction. *Circulation* **92**, 1084–1088 (1995).
41. Navab, M. *et al.* High density associated enzymes: their role in vascular biology. *Curr. Opin. Lipidol.* **9**, 449–456 (1998).
42. Schmidt, A. M., Yan, S. D., Wautier, J. L. & Stern, D. Activation of receptor for advanced glycation end products: a mechanism for chronic vascular dysfunction in diabetic vasculopathy and atherosclerosis. *Circ. Res.* **84**, 489–497 (1999).
43. Aikawa, M. *et al.* An HMG-CoA reductase inhibitor, cerivastatin, suppresses growth of macrophages expressing matrix metalloproteinases and tissue factor in vivo and in vitro. *Circulation* **103**, 276–283 (2001).
44. Ridker, P. M. *et al.* Inflammation, pravastatin, and the risk of coronary events after myocardial infarction in patients with average cholesterol levels. Cholesterol and Recurrent Events (CARE) Investigators. *Circulation* **98**, 839–844 (1998).
45. Yusuf, S. *et al.* Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. The Heart Outcomes Prevention Evaluation Study Investigators. *N. Engl. J. Med.* **342**, 145–153 (2000).
46. Dahlof, B. *et al.* Cardiovascular morbidity and mortality in the Losartan Intervention For Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. *Lancet* **359**, 995–1003 (2002).
47. Libby, P. Current concepts of the pathogenesis of the acute coronary syndromes. *Circulation* **104**, 365–372 (2001).
48. Marx, N., Sukhova, G. K., Collins, T., Libby, P. & Plutsky, J. PPAR α activators inhibit cytokine-induced vascular cell adhesion molecule-1 expression in human endothelial cells. *Circulation* **99**, 3125–3131 (1999).
49. Marx, N. *et al.* PPAR α activators inhibit tissue factor expression and activity in human monocytes. *Circulation* **103**, 213–219 (2001).
50. Neve, B. P. *et al.* PPAR α agonists inhibit tissue factor expression in human monocytes and macrophages. *Circulation* **103**, 207–212 (2001).
51. Delerive, P. *et al.* Peroxisome proliferator-activated receptor α negatively regulates the vascular inflammatory gene response by negative cross-talk with transcription factors NF- κ B and AP-1. *J. Biol. Chem.* **274**, 32048–32054 (1999).

Mast cells in autoimmune disease

Christophe Benoist & Diane Mathis

Section on Immunology and Immunogenetics, Joslin Diabetes Center; Department of Medicine, Brigham and Women's Hospital; Harvard Medical School, One Joslin Place, Boston, Massachusetts 02215, USA (e-mail: cbdm@joslin.harvard.edu)

Mast cells are known to be the primary responders in allergic reactions, orchestrating strong responses to minute amounts of allergens. Several recent observations indicate that they may also have a key role in coordinating the early phases of autoimmune diseases, particularly those involving auto-antibodies.

In imperial times, the Great Wall of China was easily breached and was not in itself a very effective defence against resolute adversaries. Rather, it was a communication route and housed, far from the imperial centre, a string of lonely guards who quickly engaged invaders and slowed their progress, while alerting and beckoning more substantial back-up forces.

Mast cells, which are scattered in skin and mucosa, have been considered in a similar outward-looking perspective^{1,2}. They are the lead effector cells in the immediate responses that can occur when sensitized individuals contact allergen through outer body surfaces. On a more beneficial note, their importance in early responses to bacterial or parasitic pathogens has become recognized in recent years. In both situations, mast cells also follow up by recruiting larger cohorts of neutrophils and lymphocytes. Recent studies suggest, however, that this picture may be incomplete and illustrate how mast cells are important in the complex cellular chains that lead to autoimmune disease.

Ehrlich's "gorged cells"

Mast cells, whose differentiation pathways and heterogeneity are still poorly understood, originate from precursors of the haematopoietic lineage and circulate in blood and the lymphatic system before homing to tissues and acquiring their final effector characteristics. The expansion, homing and maturation of mast cell precursors are influenced by several cytokines including interleukin 4 (IL-4), IL-9 and nerve growth factor (NGF)³, but stem-cell factor (SCF) binding to its receptor c-Kit seems to be the main drive for their differentiation and survival: SCF-deficient (Sl/Sl⁰) and c-Kit-deficient (W/W^v) mice are largely, albeit not completely, devoid of mast cells (for review, see refs 2, 3).

Mast cells produce an impressively broad array of mediators and cell–cell signalling molecules, and it may be this very breadth that confers on the mast cell its individuality in the immune system. Many of these mediators, including histamine, numerous specific proteases (members of the tryptase and chymase families) and tumour-necrosis factor- α (TNF- α), are released by triggered exocytosis from rich intracellular stores. The fast release of TNF- α is noteworthy because of the pleiotropic pro-inflammatory effects of this cytokine, and because mast cell granules are a plentiful source of rapidly mobilizable TNF- α (ref. 4), whose usually slower induction is the result of activated synthesis in other cell systems.

On activation, mast cells also rapidly synthesize bioactive metabolites of arachidonic acid, prostaglandins and leukotrienes. A specific program of gene expression is also activated, leading to *de novo* synthesis of several cytokines (IL-3, IL-4, IL-5, IL-6, IL-10, IL-13, IL-14 and NGF), chemokines (macrophage inflammatory protein 1 α , mono-

cyte chemoattractant protein 1 (MCP-1) and lymphotactin) and, again, TNF- α . This second-wave response comes after the immediate hypersensitivity reactions, which it amplifies. It may also bias the type of secondary events, for example, by moulding the anti-inflammatory T helper 2 (T_H2) bias of T cells in the local response to airway allergens in asthma⁵. Thus, activated mast cells signal to the vascular system through the potent vasoactivity of histamine and arachidonic metabolites, to monocytes and lymphocytes through the chemotactic and differential properties of cytokines and chemokines, and to the connective substratum through the extracellular proteases. (This is an oversimplification, however, because there is crosstalk between the different mediators and pathways, for example, in the immunomodulatory properties of prostaglandins.)

Several triggers can elicit these responses. The best characterized are allergens complexed to immunoglobulin- ϵ (IgE) molecules⁶. Because of the unusually high affinity (10⁻¹⁰ M) of the Fc receptor (FcR) for IgE (Fc ϵ R), mast cells are constantly coated with antigen-specific IgE and are, in essence, masquerading as cells of the adaptive immune system. The crosslinking of these surface-bound IgE by antigen leads to activation and degranulation. Other members of the FcR family are also active, in particular the Fc γ RIII receptor (refs 7–9). Anaphylatoxins generated by activation of the complement pathway are also potent activators of some mast cells^{10,11}. Bacterial microbes can trigger mast cells through Toll-like receptors (TLRs), endowing them with the broad 'pattern recognition' capability of the TLR system, which is probably an important element of their antibacterial responses^{12,13}. Some cytokines and chemokines activate mast cells, in particular TNF- α and MCP-1, which are themselves released by mast cells, thus raising the potential for a positive feedback loop. Finally, activation of mast cells by co-culture with activated T cells has been described, but it is not clear what molecular mediators may be involved^{14,15}. Direct crosstalk by surface molecules on T cells and mast cells may be important in this context.

Autoimmune disease in the brain

The recent spark of interest in a role for mast cells in initiating or propagating autoimmune disease was prompted by studies on multiple sclerosis and its animal model, experimental allergic encephalomyelitis (EAE)¹⁶. Multiple sclerosis is a chronic inflammatory disorder of the central nervous system (CNS), which is characterized by a breach of the blood–brain barrier, mononuclear cell infiltration of white matter and eventual demyelination. A similar autoimmune disease can be induced in susceptible rodent strains by injecting different myelin components, including myelin basic protein (MBP), proteolipid protein and myelin oligodendrocyte glycoprotein (MOG).

Both multiple sclerosis and EAE depend critically on pro-inflammatory T helper 1 (T_H1) $CD4^+$ T cells, B cells, and more specifically the antibodies that they produce, may also be important, although this is still under debate. Numerous studies, dating as far back as 100 years, have reported a correlation between the number and/or distribution of mast cells and the development of multiple sclerosis or EAE (reviewed in ref. 17). Evidence of mast cell activation in the course of the disease came from the demonstration of increasing degranulation¹⁸ and increased amounts of proteolytic enzymes such as trypsin in cerebrospinal fluid¹⁹. In addition, drugs considered to 'stabilize' mast cells (for example, cromolyn sodium) have been shown to ameliorate the severity of EAE^{20–22}.

Although these observations were highly suggestive of an essential role for mast cells in these CNS autoimmune diseases, the association remained indirect until the recent studies of Brown and colleagues²³. These researchers showed that mice lacking mast cells (W/W^v mice) develop EAE later and less severely than do control mice in response to injection of MOG. Complementation of W/W^v mice with immature mast cells derived *in vitro* restores typical EAE susceptibility. Mast cell function seems to be the result of binding antibodies, as it was found to be dependent on expression, by the mast cells, of the $Fc\gamma R^{17}$. Notably, Brown and colleagues¹⁷ subsequently showed that their procedure does not result in reconstitution of mast cells in CNS tissues, suggesting that mast cells might be exerting their crucial influence outside the inflammatory lesion.

Another line of evidence has independently piqued interest in a role for mast cells in multiple sclerosis and EAE. Gene expression profiling of multiple sclerosis brain lesions detected an unexpectedly high contribution of transcripts either derived from mast cells or otherwise associated with the allergic response, including transcripts encoding histamine receptors, proteases and other inflammatory mediators^{24,25}. These findings rekindle interest in the perplexing finding that the transfer of MBP-specific T_H2 cells to healthy recipients unexpectedly provoked a variant form of EAE characterized by eosinophilic infiltrates into the CNS²⁶.

Autoimmune disease in the joint

A potential role for mast cells in rheumatoid arthritis has also been highlighted recently. Rheumatoid arthritis is a chronic inflammatory disease of the diarthrodial joints. K/BxN mice spontaneously develop a joint disorder that has many similarities to rheumatoid arthritis²⁷. Although the development of disease in this model is initiated by T cells, it also requires B cells, and immunoglobulin- γ (IgG) antibodies from an arthritic donor can induce disease in a healthy host. The target of both the pathogenic T cells and arthritogenic antibodies is the ubiquitous cytoplasmic enzyme glucose-6-phosphate isomerase (GPI)²⁸. This enzyme and antibodies against it aggregate as immune complexes at the surface of the articular cavity, where they initiate an inflammatory cascade involving the alternative pathway of complement (acting through C5a), FcRs (in particular, $Fc\gamma RIII$), neutrophils and cytokines such as IL-1 and TNF- α (refs 29–31).

Now it seems that mast cells are also important in this disease process³². Both Sl/Sl^d and W/W^v mice are resistant to the induction of arthritis by antibodies against GPI. More definitively, reconstitution of these mice with mast cell precursors restores sensitivity to disease induction. Notably, one of the first events detected after injection of arthritogenic antibodies into wild-type mice is mast cell degranulation in the joint but not in other tissues. This very early event is already apparent an hour after antibody administration, before the recruitment of neutrophils. These results prompted the conclusion that mast cells might have an early, coordinating role in this model of rheumatoid arthritis.

The generality of this conclusion is supported by observations from other murine models of rheumatoid arthritis and from individuals affected with rheumatoid arthritis. Mast cells accumulate in the swollen paws of mice suffering from collagen-induced arthritis, and they degranulate during the disease process³³. Salbutamol is a

β_2 -adrenergic agonist that prevents mast cell degranulation, and this drug had a strong therapeutic effect on the progression of collagen-induced arthritis³³. Mast cell deficiency was also found to inhibit the course of antigen-induced arthritis in mice, although the effect was rather mild³⁴. Mast cells also accumulate in the synovial tissues and fluids of humans suffering from rheumatoid arthritis^{35,36}, reflecting the presence of mast cell chemotactic or survival activities such as SCF and transforming growth factor- β in the synovial fluid³⁷. The invading mast cells produce several inflammatory mediators, notably TNF- α , IL-1 β and vascular endothelial growth factor (VEGF)^{35,38}. Notably, TNF- α can induce further production of SCF by synovial fibroblasts, potentially augmenting mast cell recruitment and thereby creating an amplification loop.

Autoimmune disease in the skin

Bullous pemphigoid seems to present a situation that is highly similar to the one that unfolds in K/BxN mice. This autoimmune skin disease is characterized by subepidermal blisters resulting from auto-antibodies against two hemidesmosomal antigens, BP230 and BP180 (ref. 39). The key features of the human disease can be mimicked by injecting neonatal mice intradermally with IgG antibodies directed against murine BP180 (ref. 40). The antibody-induced disease has been known for some time to require activation of the complement pathway⁴¹ and the accumulation of neutrophils⁴². Recently, it has been also shown to depend critically on mast cells⁴³.

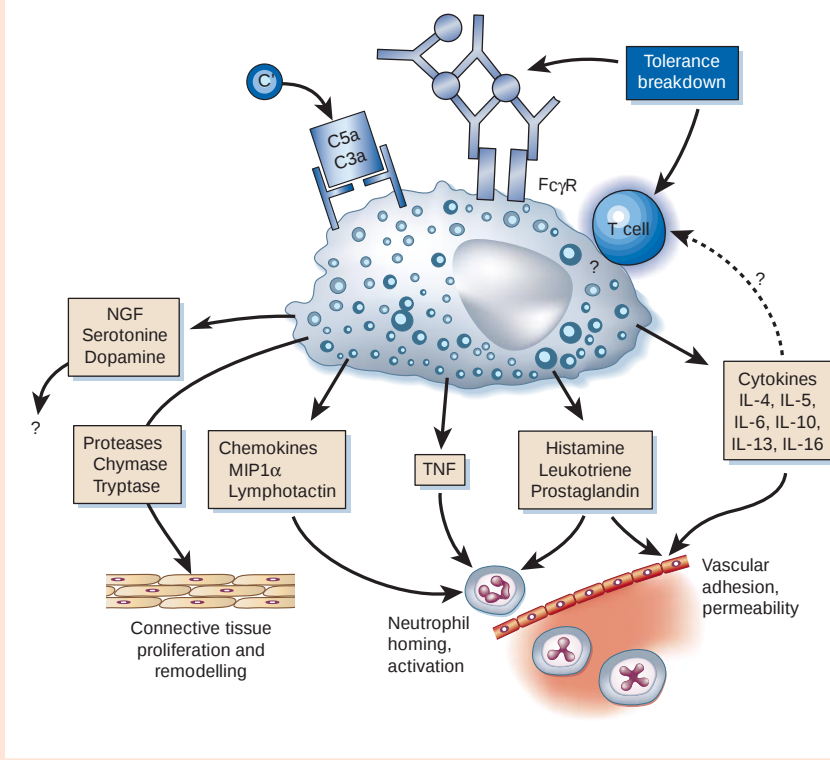
Mast cell degranulation was one of the first responses detected after the injection of antibodies against BP180, occurring only 1 h after administration and preceding neutrophil accumulation and skin blistering⁴³. Injection of antibodies against BP180 into mice lacking mast cells (W/W^v or Sl/Sl^d) did not induce bullous pemphigoid, nor did their injection into wild-type mice pre-treated with cromolyn sodium. But mice lacking mast cells that were reconstituted intradermally with mast cells derived *in vitro* showed typical features of disease. In the absence of mast cells, IgG still accumulated in the skin and the complement pathway was activated to yield C3a and C5a, but neutrophils were no longer recruited to the dermal lesion. Bullous pemphigoid could be induced in mast-cell-deficient mice injected with antibodies against BP180 if neutrophils or the potent neutrophil attractant IL-8 were injected intradermally. Thus, it was concluded that the crucial role of mast cells in murine bullous pemphigoid is to recruit neutrophils to the developing lesion. A similar process might also occur in the human disease, because degranulated mast cells are a prominent feature of the skin blisters of individuals affected with bullous pemphigoid⁴⁴, and mast-cell-derived chemoattractants are present at high concentrations in blister fluids^{45,46}.

There are several other examples of autoimmune disorders in which mast cells have been implicated, although often only by 'guilt by association'. These include Sjogren's syndrome⁴⁷, chronic idiopathic urticaria⁴⁸, thyroid eye disease⁴⁹ and experimental vasculitis⁵⁰. For these disorders it will be important to provide evidence, as in the three diseases highlighted here, that mast cells are more than bystanders that become activated in the inflammatory maelstrom and are involved directly in the complex chain of cellular events that lead to autoimmune damage.

The role of mast cells

Where, however, are mast cells positioned in this chain? What triggers them into action, and which are the important relay molecules (Fig. 1)? For the antibody-mediated models (pemphigoid and K/BxN arthritis), there is no dearth of candidates that might activate mast cells: the two main consequences of immune complex formation — the production of complement-derived anaphylatoxins and $Fc\gamma R$ crosslinking — can both trigger mast cells efficiently^{7–11}. It will be important to pinpoint which of these pathways is involved by analysing mast cell degranulation in knockout animals and by reconstituting W/W^v mice with mast cells derived from complement- or FcR -deficient mice.

Figure 1 The mast cell as an integrator or amplifier of autoimmune responses. The breakdown of tolerance and/or immunoregulatory mechanisms leads to autoimmune activation and recognition in the tissues. These responses, which are 'adaptive' in their anti-self specificity, generate primary 'innate' inputs into mast cells, such as immune complex binding to FcRs, and C3a and C5a anaphylatoxins of the complement pathway binding to specific receptors. The molecular route for direct 'bystander' activation of mast cells by T cells remains conjectural. The mast cell, owing to the abundance and diversity of secondary mediators in its granules, responds by activating a host of pathways, thus amplifying the local response. Vascular permeability is increased, allowing influx of additional molecules (antibody, complement). The adhesiveness of the vascular endothelium is increased, facilitating the homing of leukocytes (and in particular neutrophils) provoked by chemokine and TNF- α release. These leukocytes are also activated by the same cytokines. Mast cell mediators may be also involved in remodelling connective tissue, or in biasing secondary T-cell responses. Mast cell activation may also signal to local neuronal constituents by the release of NGF, serotonin or dopamine. Thus, the mast cell takes in what may be a low pro-inflammatory input and amplifies it to bring about a much wider response.



For the EAE models, in which T cells are classically thought to be the effectors, one might have invoked the effect that activated T cells have on mast cells^{14,15}. But the effectiveness of mast cell reconstitution seems to be dependent on the presence of Fc γ R¹⁷, pointing to an involvement of antibodies against MOG in this disease. Notably, MOG-induced EAE is the model that is thought to be most dependent on antibodies for lesion development; thus, here again the mast cell contribution may be antibody-dependent. These data do not rule out a direct interaction between T cells and mast cells, and it will be interesting to examine the role of mast cells in 'pure' T-cell-mediated autoimmune diseases, such as diabetes.

The heterogeneity of mast cell populations, their variations in different tissue environments and how they may differentially integrate input from different stimuli are incompletely understood facets of their biology. Is the response of an airway mast cell to an allergen that crosslinks IgE receptors the same as that of a joint mast cell to deposited IgG? Complex interactions take place between the intracellular signals elicited when Fc γ R and Fc ϵ R are both engaged, and these influence the mediators that are released or induced^{7,9}. It will be important to determine how concomitant triggering of mast cells through the Fc γ R, C5a and other secondary byproducts of immune complexes may be integrated differentially by mast cells, thereby leading to consequences as different as a pemphigus blister or an EAE plaque.

Downstream of mast cell activation, all of the events described in IgE-induced allergic responses^{1,2} have the potential to fan the autoimmune flames. For example, there will be increased permeability of the local vasculature, which will recruit even more immune complexes into the lesion; notably, local oedema is one of the earliest events in the unfolding of antibody-induced arthritis. There will be modifications of vascular adhesive properties contributing to the recruitment of leukocytes by chemokines, comparable to the mast-cell-mediated influx of neutrophils in models of peritonitis^{11,51,52}. In the arthritis model, neutrophils are also essential³⁰, and it may be that the sequential mast cell/neutrophil tandem will constitute a frequently recurring theme. The very early timing of mast cell

degranulation in both the bullous pemphigoid and rheumatoid arthritis mouse models are consistent with that view. In the peritonitis models, TNF- α seems to be the essential mediator for neutrophil recruitment^{51,52}. Given the central role that TNF- α seems to have in arthritis, it will be interesting to see whether it is also the principal contribution of the mast cell.

In both asthma and arthritis, the worst damage lies not so much in the immediate inflammation as in the subsequent tissue reorganization and chronic inflammation. Connective tissue proliferation leads to loss of organ function, whether as an eroding pannus in the joint or as thickened and hyperreactive bronchi. Arthritis, in particular, has been described as a tumour-like anarchic proliferation of synovio-cytes. Several mast cell products have strong trophic effects, including classical growth factors (NGF, epidermal growth factor, VEGF), but some of the mast cell proteases also have mitogenic properties². One might propose that mast cells are important contributors in the anarchic joint reconstruction triggered by the autoimmune attack. Last, as suggested by Brown and colleagues^{17,23}, there is the intriguing possibility that mast cell activation also feeds back to the initiating autoimmune responses in lymphocytes. The release of tissue neo-antigens through proteolysis might contribute to the epitope spreading observed in EAE. Or, as in asthma, the locally released cytokines might bias T-cell phenotypes, enhancing a T_H2 response that would bolster the dangerous production of auto-antibodies.

Autoimmune diseases such as multiple sclerosis or rheumatoid arthritis are complex and involve long and convoluted molecular and cellular chains, with many possible points for therapeutic intervention. Yet the demonstration of an obligate passage through mast cells in these animal models opens the perspective of harnessing agents that modulate mast cell homeostasis or function to treat human disease.

Mast cells have been positioned historically in the private domain of allergists and have been largely ignored by the autoimmunity field. This ignorance can no longer be sustained as the demarcation between autoimmunity and allergy becomes fuzzy. This is illustrated by the anaphylactic reactions induced, under certain conditions, by

injecting myelin proteins or peptides into mice or individuals with multiple sclerosis^{53–55}. And the view of mast cells as a ring of outward-looking sentinels can no longer hold. Their scope clearly includes the inner realm as well. □

doi:10.1038/nature01324

1. Galli, S. J., Maurer, M. & Lantz, C. S. Mast cells as sentinels of innate immunity. *Curr. Opin. Immunol.* **11**, 53–59 (1999).
2. Mekori, Y. A. & Metcalfe, D. D. Mast cells in innate immunity. *Immunol. Rev.* **173**, 131–140 (2000).
3. Galli, S. J., Zsebo, K. M. & Geissler, E. N. The kit ligand, stem cell factor. *Adv. Immunol.* **55**, 1–96 (1994).
4. Young, J. D., Liu, C. C., Butler, G., Cohn, Z. A. & Galli, S. J. Identification, purification, and characterization of a mast cell-associated cytolytic factor related to tumor necrosis factor. *Proc. Natl Acad. Sci. USA* **84**, 9175–9179 (1987).
5. Williams, C. M. & Galli, S. J. The diverse potential effector and immunoregulatory roles of mast cells in allergic disease. *J. Allergy Clin. Immunol.* **105**, 847–859 (2000).
6. Turner, H. & Kinet, J. P. Signalling through the high-affinity IgE receptor FcεRI. *Nature* **402**, B24–B30 (1999).
7. Daeron, M., Malbec, O., Latour, S., Arock, M. & Fridman, W. H. Regulation of high-affinity IgE receptor-mediated mast cell activation by murine low-affinity IgG receptors. *J. Clin. Invest.* **95**, 577–585 (1995).
8. Sylvestre, D. L. & Ravetch, J. V. A dominant role for mast cell Fc receptors in the arthus reaction. *Immunity* **5**, 387–390 (1996).
9. Okayama, Y., Hagaman, D. D. & Metcalfe, D. D. A comparison of mediators released or generated by IFN-γ-treated human mast cells following aggregation of Fc gamma RI or Fc epsilon RI. *J. Immunol.* **166**, 4705–4712 (2001).
10. Austen, K. F. & Becker, E. L. Mechanisms of immunologic injury of rat peritoneal mast cells. II. Complement requirement and phosphonate ester inhibition of release of histamine by rabbit anti-rat γ-globulin. *J. Exp. Med.* **124**, 397–416 (1966).
11. Prodeus, A. P., Zhou, X., Maurer, M., Galli, S. J. & Carroll, M. C. Impaired mast cell-dependent natural immunity in complement C3-deficient mice. *Nature* **390**, 172–175 (1997).
12. Supajatura, V. *et al.* Protective roles of mast cells against enterobacterial infection are mediated by Toll-like receptor 4. *J. Immunol.* **167**, 2250–2256 (2001).
13. Applequist, S. E., Wallin, R. P. & Ljunggren, H. G. Variable expression of Toll-like receptor in murine innate and adaptive immune cell lines. *Int. Immunol.* **14**, 1065–1074 (2002).
14. Schmitt, E., Huls, C., Nagel, B. & Rude, E. Characterization of a T-cell-derived mast cell costimulatory activity (MCA) that acts synergistically with interleukin 3 and interleukin 4 on the growth of murine mast cells. *Cytokine* **2**, 407–415 (1990).
15. Mekori, Y. A. & Metcalfe, D. D. Mast cell-T cell interactions. *J. Allergy Clin. Immunol.* **104**, 517–523 (1999).
16. Steinman, L. Multiple sclerosis: a two-stage disease. *Nature Immunol.* **2**, 762–764 (2001).
17. Brown, M., Tanzola, M. & Robbie-Ryan, M. Mechanisms underlying mast cell influence on EAE disease course. *Mol. Immunol.* **38**, 1373 (2002).
18. Brenner, T., Soffer, D., Shalit, M. & Levi-Schaffer, F. Mast cells in experimental allergic encephalomyelitis: characterization, distribution in the CNS and *in vitro* activation by myelin basic protein and neuropeptides. *J. Neurol. Sci.* **122**, 210–213 (1994).
19. Rozniecki, J. J., Hauser, S. L., Stein, M., Lincoln, R. & Theoharides, T. C. Elevated mast cell tryptase in cerebrospinal fluid of multiple sclerosis patients. *Ann Neurol.* **37**, 63–66 (1995).
20. Brosnan, C. F. & Tansey, F. A. Delayed onset of experimental allergic neuritis in rats treated with reserpine. *J. Neuropathol. Exp. Neurol.* **43**, 84–93 (1984).
21. Dietsch, G. N. & Hinrichs, D. J. The role of mast cells in the elicitation of experimental allergic encephalomyelitis. *J. Immunol.* **142**, 1476–1481 (1989).
22. Seelard, P. A., Yasui, D., Weiner, H. L. & Johnson, D. Treatment of experimental allergic neuritis with nedocromil sodium. *J. Neuroimmunol.* **25**, 221–226 (1989).
23. Secor, V. H., Secor, W. E., Gutekunst, C. A. & Brown, M. A. Mast cells are essential for early onset and severe disease in a murine model of multiple sclerosis. *J. Exp. Med.* **191**, 813–822 (2000).
24. Lock, C. *et al.* Gene-microarray analysis of multiple sclerosis lesions yields new targets validated in autoimmune encephalomyelitis. *Nature Med.* **8**, 500–508 (2002).
25. Chabas, D. *et al.* The influence of the proinflammatory cytokine, osteopontin, on autoimmune demyelinating disease. *Science* **294**, 1731–1735 (2001).
26. Lafaille, J. J. *et al.* Myelin basic protein-specific T helper 2 (Th2) cells cause experimental autoimmune encephalomyelitis in immunodeficient hosts rather than protect them from the disease. *J. Exp. Med.* **186**, 307–312 (1997).
27. Kouskoff, V. *et al.* Organ-specific disease provoked by systemic autoreactivity. *Cell* **87**, 811–822 (1996).
28. Matsumoto, I., Staub, A., Benoist, C. & Mathis, D. Arthritis provoked by linked T and B cell recognition of a glycolytic enzyme. *Science* **286**, 1732–1735 (1999).
29. Ji, H. *et al.* Arthritis critically dependent on innate immune system players. *Immunity* **16**, 157–168 (2002).
30. Wipke, B. T. & Allen, P. M. Essential role of neutrophils in the initiation and progression of a murine model of rheumatoid arthritis. *J. Immunol.* **167**, 1601–1608 (2001).
31. Ji, H. *et al.* Critical roles for interleukin-1 and tumor necrosis factor-α in antibody-induced arthritis. *J. Exp. Med.* **196**, 77–85 (2002).
32. Lee, D. M. *et al.* Mast cells: a cellular link between autoantibodies and inflammatory arthritis. *Science* **297**, 1689–1692 (2002).
33. Malfait, A. M. *et al.* The beta2-adrenergic agonist salbutamol is a potent suppressor of established collagen-induced arthritis: mechanisms of action. *J. Immunol.* **162**, 6278–6283 (1999).
34. van den Broek, M. F., van den Berg, W. B. & Van de Putte, L. B. The role of mast cells in antigen induced arthritis in mice. *J. Rheumatol.* **15**, 544–551 (1988).
35. Woolley, D. E. & Tetlow, L. C. Mast cell activation and its relation to proinflammatory cytokine production in the rheumatoid lesion. *Arthritis Res.* **2**, 65–74 (2000).
36. Crisp, A. J., Chapman, C. M., Kirkham, S. E., Schiller, A. L. & Krane, S. M. Articular mastocytosis in rheumatoid arthritis. *Arthritis Rheum.* **27**, 845–851 (1984).
37. Olsson, N., Ulfgren, A. K. & Nilsson, G. Demonstration of mast cell chemotactic activity in synovial fluid from rheumatoid patients. *Ann. Rheum. Dis.* **60**, 187–193 (2001).
38. Yamada, T. *et al.* Localization of vascular endothelial growth factor in synovial membrane mast cells: examination with “multi-labelling subtraction immunostaining”. *Virchows Arch.* **433**, 567–570 (1998).
39. Stanley, J. R. in *Fitzpatrick's Dermatology in General Medicine* (eds Freedberg, I. M. *et al.*) 666–671 (McGraw-Hill, New York, 1999).
40. Liu, Z. *et al.* A passive transfer model of the organ-specific autoimmune disease, bullous pemphigoid, using antibodies generated against the hemidesmosomal antigen, BP180. *J. Clin. Invest.* **92**, 2480–2488 (1993).
41. Liu, Z. *et al.* The role of complement in experimental bullous pemphigoid. *J. Clin. Invest.* **95**, 1539–1544 (1995).
42. Liu, Z. *et al.* A major role for neutrophils in experimental bullous pemphigoid. *J. Clin. Invest.* **100**, 1256–1263 (1997).
43. Chen, R. *et al.* Mast cells play a key role in neutrophil recruitment in experimental bullous pemphigoid. *J. Clin. Invest.* **108**, 1151–1158 (2001).
44. Wintrob, B. U., Mihm, M. C., Jr, Goetzl, E. J., Soter, N. A. & Austen, K. F. Morphologic and functional evidence for release of mast-cell products in bullous pemphigoid. *N. Engl. J. Med.* **298**, 417–421 (1978).
45. Baba, T. *et al.* An eosinophil chemotactic factor present in blister fluids of bullous pemphigoid patients. *J. Immunol.* **116**, 112–116 (1976).
46. Katayama, I., Doi, T. & Nishioka, K. High histamine level in the blister fluid of bullous pemphigoid. *Arch. Dermatol. Res.* **276**, 126–127 (1984).
47. Kontinen, Y. T. *et al.* Mast cell derangement in salivary glands in patients with Sjogren's syndrome. *Rheumatol. Int.* **19**, 141–147 (2000).
48. Napoli, D. C. & Freeman, T. M. Autoimmunity in chronic urticaria and urticarial vasculitis. *Curr Allergy Asthma Rep.* **1**, 329–336 (2001).
49. Ludgate, M. & Baker, G. Unlocking the immunological mechanisms of orbital inflammation in thyroid eye disease. *Clin. Exp. Immunol.* **127**, 193–198 (2002).
50. Kiely, P. D., Pecht, I. & Oliveira, D. B. Mercuric chloride-induced vasculitis in the Brown Norway rat: αβ T cell-dependent and -independent phases: role of the mast cell. *J. Immunol.* **159**, 5100–5106 (1997).
51. Zhang, Y., Ramos, B. F. & Jakschik, B. A. Neutrophil recruitment by tumor necrosis factor from mast cells in immune complex peritonitis. *Science* **258**, 1957–1959 (1992).
52. Malaviya, R., Ikeda, T., Ross, E. & Abraham, S. N. Mast cell modulation of neutrophil influx and bacterial clearance at sites of infection through TNF-α. *Nature* **381**, 77–80 (1996).
53. Pedotti, R. *et al.* An unexpected version of horror autotoxicus: anaphylactic shock to a self-peptide. *Nature Immunol.* **2**, 216–222 (2001).
54. Bielekova, B. *et al.* Encephalitogenic potential of the myelin basic protein peptide (amino acids 83–99) in multiple sclerosis: results of a phase II clinical trial with an altered peptide ligand. *Nature Med.* **6**, 1167–1175 (2000).
55. Kappos, L. *et al.* Induction of a non-encephalitogenic type 2 T helper-cell autoimmune response in multiple sclerosis after administration of an altered peptide ligand in a placebo-controlled, randomized phase II trial. The Altered Peptide Ligand in Relapsing MS Study Group. *Nature Med.* **6**, 1176–1182 (2000).

Inflammation and therapeutic vaccination in CNS diseases

Howard L. Weiner & Dennis J. Selkoe

Center for Neurologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, Massachusetts 02115, USA

The spectrum of inflammatory diseases of the central nervous system has been steadily expanding from classical autoimmune disorders such as multiple sclerosis to far more diverse diseases. Evidence now suggests that syndromes such as Alzheimer's disease and stroke have important inflammatory and immune components and may be amenable to treatment by anti-inflammatory and immunotherapeutic approaches. The notion of 'vaccinating' individuals against a neurodegenerative disorder such as Alzheimer's disease is a marked departure from classical thinking about mechanism and treatment, and yet therapeutic vaccines for both Alzheimer's disease and multiple sclerosis have been validated in animal models and are in the clinic. Such approaches, however, have the potential to induce unwanted inflammatory responses as well as to provide benefit.

Inflammation of the central nervous system (CNS) may be the result of both innate and adaptive immune responses. In Alzheimer's disease (AD) an innate immune response is triggered by local production of amyloid β -protein (A β), whereas in multiple sclerosis (MS) an adaptive immune response directed against myelin components initiates inflammation in the CNS (Table 1). Adaptive immune responses involving antibody- or cell-mediated responses have differential effects in AD and MS, and in animal models of these diseases. In addition, the recent appearance of encephalitis in individuals with AD that have been immunized with A β has parallels to underlying mechanisms of cell-mediated adaptive immune responses in MS, in which pro-inflammatory T-cell responses seem to drive the disease. These features of inflammation, which are outlined for AD and MS in Table 1, are reviewed here in terms of both disease pathogenesis and therapy.

Multiple sclerosis

Multiple sclerosis is an inflammatory disease of the central nervous system characterized by perivascular cuffs of mononuclear cells that include both lymphocytes and macrophages¹. This infiltration leads to damage of the myelin sheath and the underlying axon. Activation of microglia and astrocytes occurs in MS, but it is secondary to infiltrating lymphocytes. In the initial stages of the disease, the inflammation that occurs in MS is episodic and associated with discrete attacks of neurological dysfunction followed by recovery, which may leave residual neurological damage. Subsequently the disease often becomes more progressive, developing to a stage where there is less

inflammation and nervous system damage is caused by a degenerative process initiated by the inflammation.

The episodic inflammation that is classic of MS is clearly visualized by magnetic resonance imaging (MRI) scans of the brain after administration of the contrast material gadolinium². Gadolinium crosses an open blood-brain barrier created by the inflammation and highlights discrete areas of inflammation. The duration of enhanced inflammation in individuals receiving weekly MRI scans is 4–8 weeks, and virtually all new lesions enhance in their earliest phases³. When the acute inflammation resolves, it leaves a scar and tissue damage. This can be seen in the three-dimensional MRI images in Fig. 1, which were recorded over a 1-yr period in a single individual affected with MS. The new inflammatory focus can be seen appearing adjacent to the ventricle and then beginning to resolve. The inflammatory process of MS is associated with a complex cascade of inflammatory molecules and mediators, including chemokines, adhesion molecules associated with activated endothelial cell walls and matrix metalloproteases^{4–6}.

The cause of the recurrent inflammation in MS is now generally accepted to be autoimmune in nature, that is, a cell-mediated autoimmune attack against the white matter sheath⁷. An alternative explanation for the episodic and chronic inflammation that is the hallmark of MS is the presence of a virus or infectious agent that has persistently infected the nervous system. But although infectious agents have been extensively sought in MS, none has been isolated⁸. Viruses and infectious agents are, however, thought to be important in triggering the immune system and the immune attack on the nervous system⁹. Given the inflammatory nature of the pathological process and the

Figure 1 Three-dimensional MRI scans of multiple sclerosis showing old scars (yellow), a new area of inflammation that appears on day 266 (red) and evidence of resolution of the inflammation on day 362. The MS lesions are shown in relation to the ventricular cavities of the brain (blue).



Table 1 Immune responses and CNS inflammation in Alzheimer's disease and multiple sclerosis

Immune response	Features	AD/APP mouse model	MS/EAE mouse model
Innate*	Microglia/astrocyte activation	Induced by A β_{1-42}	Induced by myelin-reactive T cells
Adaptive (humoral)	Antibody	Antibodies against A β are therapeutic in mouse model	Antibodies against MOG worsen EAE†
Adaptive (cell-mediated)‡			
Pro-inflammatory T-cell responses (T _H 1)	IFN- γ (IL-12)	Meningoencephalitis after A β immunization in individuals with AD; encephalitis in mice	Anti-myelin T _H 1 responses initiate MS and EAE
Anti-inflammatory T-cell responses (T _H 2, T _H 3, T _R 1, CD25 ⁺ cells)	IL-4, TGF- β , IL-10	Unknown in AD; induced by nasal A β in mouse model	Protective in MS and EAE

*The innate immune response in AD and the APP mouse model of AD involves the activation of microglia and astrocytes. This activation is induced by the accumulation of A β_{1-42} in the CNS. Microglia and astrocyte activation also occurs in MS but is induced by myelin-reactive T cells that are activated in the periphery and migrate to the CNS.

†The adaptive humoral immune response after immunization with A β has marked beneficial effects in the APP mouse model of AD, leading to the clearance of A β from the brain and improvement in cognitive function. There is no significant endogenous adaptive humoral response against A β in individuals with AD, and the effect of exogenously induced antibodies against A β in AD awaits clinical trials. Adaptive endogenous humoral immune responses against myelin antigens (such as MBP and MOG) have been identified in individuals affected with MS but are not thought to have a central role in the disease. Antibodies against MOG worsen the T-cell-mediated EAE model of MS.

‡A pro-inflammatory (T_H1)-type adaptive cell-mediated immune response against A β is postulated to be the cause of meningoencephalitis in individuals with AD who are vaccinated with A β in adjuvant.

Endogenous adaptive T-cell responses (pro- or anti-inflammatory) against A β have not been identified in these individuals although continuing investigations suggest that they may exist. T_H1-type cell-mediated adaptive immune responses against myelin antigens have been shown to cause EAE, and accumulating evidence suggests that they have a central role in MS. Treatments that decrease T_H1-type responses and induce anti-inflammatory (IL-4, TGF- β , IL-10) responses are beneficial in MS and protective in EAE. Note that in immune-deficient mice T_H2-type responses can induce a form of EAE.

autoimmune hypothesis, one might expect that anti-inflammatory immunosuppressive drugs would reduce inflammation, as measured by MRI imaging, and positively affect the clinical course. Indeed, this has been shown clearly with agents such as mitoxantrone¹⁰, a chemotherapy drug, and cyclophosphamide^{11,12}, a chemotherapy drug that is also used in other inflammatory conditions such as lupus nephritis and inflammatory muscle disease. The most widely used drugs in MS, β -interferon and glatiramer acetate, have anti-inflammatory and immunomodulatory effects and are discussed in more detail below¹³.

Adaptive cell-mediated immune responses in MS

The adaptive immune system can be classified broadly into cellular and humoral (antibody)-type responses. Among cellular responses, different types or classes of cellular immune responses have been identified that are essential to understanding the mechanisms of the inflammatory process in MS and to devising strategies to control it. As discussed below, the different classes of cell-mediated immune response have important implications for attempts to develop a vaccination strategy not only for MS but also for AD.

Cellular immune responses can be classified as T_H1-type or T_H2-type responses (Fig. 2), depending on how they differentiate from T_H0 precursors¹⁴. T_H1 (or pro-inflammatory) responses are induced when T cells differentiate in the presence of interleukin 12 (IL-12), and T_H1 cells are characterized by the secretion of interferon- γ (IFN- γ) and inflammatory mediators such as tumour-necrosis factor- α (TNF- α). T_H1-type responses are important in fighting viral infections, and MS seems to be a cell-mediated autoimmune disease of a T_H1 type. Anti-inflammatory T-cell responses include both T_H2 responses and T cells that have been classified as 'regulatory cells'. T_H2 responses are induced when T cells differentiate in the presence of IL-4, and T_H2-type cells secrete anti-inflammatory cytokines such as IL-4 and IL-10. T_H2-type responses are important in fighting parasitic infections, and T_H1 and T_H2 responses may cross-regulate each other.

Another class of T cell comprises regulatory cells that can downregulate T_H1-type inflammatory processes. Different types of regulatory cell have been described¹⁵. T_H3 cells act primarily through the secretion of transforming growth factor- β (TGF- β) and are preferentially induced at mucosal surfaces¹⁶. T_R1 cells (T regulatory cell 1) act primarily through the secretion of IL-10 (ref. 17), and CD4⁺CD25⁺ regulatory cells are T cells that express CD25 (IL-2 receptor) and exert potent regulatory function through cell contact and also through cytokines such as IL-10 and TGF- β (ref. 18). If MS is a T_H1-type cell-mediated autoimmune disease, it might be possible to regulate the T_H1 responses by the induction of regulatory cell populations.

The induction of T_H1-type myelin-reactive cells and their migration into the nervous system is shown in Fig. 2a. It is postulated that

T_HP (T precursor) myelin-reactive T cells are induced to differentiate into myelin-reactive T_H1 cells when an antigen that crossreacts with a myelin antigen is presented to a T cell by an antigen-presenting cell in the context of IL-12 and co-stimulatory molecules. It is generally thought that viruses with structures that crossreact with myelin antigens act as crossreactive antigens¹⁹. T_H1 T cells that react with myelin antigens, such as proteolipid protein (PLP), myelin basic protein (MBP) and myelin oligodendrocyte glycoprotein (MOG), cross the blood-brain barrier where the myelin antigens are represented to the T cell by antigen-presenting cells in the brain (microglia cells), and an inflammatory cascade is triggered with the release of inflammatory mediators that cause damage to the myelin sheath and ultimately the underlying axon.

One of the primary animal models for MS, experimental allergic encephalomyelitis (EAE), is induced by immunizing different mouse or rat strains with a myelin autoantigen (such as MBP, PLP or MOG) given in complete Freund's adjuvant, which induces a T_H1-type cell-mediated response against the myelin antigen. In EAE, myelin-reactive T_H1-type CD4⁺ T cells migrate from the periphery into the CNS, where they also initiate a cascade of immune-mediated damage (Fig. 2a). In animals, EAE can be induced by the adoptive transfer of T_H1-type CD4⁺ cells specific for one of the myelin proteins.

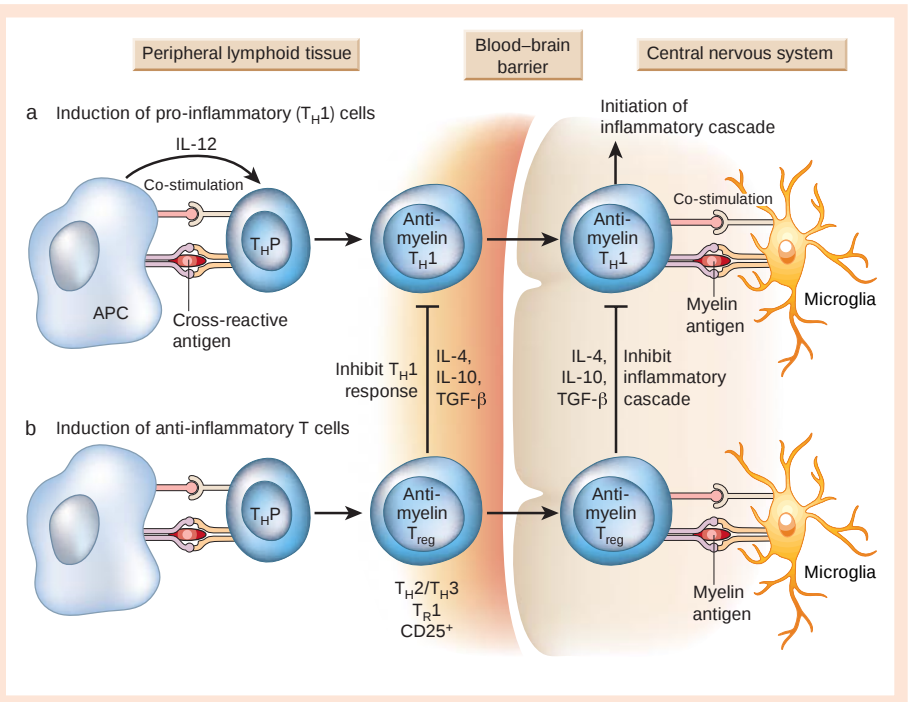
The hypothesis that MS is a inflammatory T_H1-type disease is supported by several observations. First, it has been shown directly by the effects of γ -interferon, the prototypic T_H1 cytokine, which when administered to individuals with MS caused clinical exacerbations²⁰. Second, individuals affected with MS have a T_H1 bias, as indicated by increased concentrations of IL-12 (refs 21, 22) and IL-18 (ref. 23), both of which induce IFN- γ and increase T_H1-type chemokine receptor expression^{5,24}. Last, IL-12-secreting cells in the peripheral blood are linked to inflammation in the CNS, as measured by gadolinium enhancement on MRI imaging²⁵; increased numbers of IL-12-secreting cells in the blood are associated with gadolinium enhancement, and cyclophosphamide decreases the number of IL-12-secreting cells, which is linked to clinical response²⁶.

In addition to IL-12, it has been shown recently that osteopontin is important in T_H1 differentiation in autoimmune demyelinating disease^{27,28}. The most widely used immunomodulatory drug in MS, β -interferon, seems to have two broad mechanisms of action: it decreases γ -interferon secretion by cells in the peripheral blood and blocks the migration of T cells across the blood-brain barrier¹³.

Vaccination

The term 'vaccination' stems from the original observation of Jenner and his use of subcutaneous administration of cowpox to prevent the subsequent development of smallpox. Since then, the term vaccination has acquired a broader meaning. According to current immunological theory, vaccination is no longer restricted to administering infectious agents but applies to manipulating the immune system in a

Figure 2 Inflammation and immune mechanisms in multiple sclerosis. **a**, Multiple sclerosis is thought to be induced by the generation of T_H1 -type myelin-reactive cells from precursor cells (T_HP), which are presumed to be triggered by crossreactive antigens such as viruses in the context of co-stimulatory molecules and IL-12. T_H1 -type cells directed against myelin migrate into the nervous system where they re-encounter myelin antigens presented by microglia and are restimulated to initiate a destructive inflammatory cascade. Immune therapy involves the induction of anti-inflammatory regulatory T cells (T_H2 , T_H3 , T_R1 , $CD25^+$ cells) that secrete anti-inflammatory cytokines, such as IL-4, IL-10 and TGF- β , or may also act by cell to cell contact ($CD25^+$ cells). These regulatory cells inhibit T_H1 responses in the periphery and/or migrate to the CNS, where they are re-stimulated by local microglia cells and inhibit or suppress the local inflammatory cascade in the CNS. Regulatory cells can be induced by different means, including glatiramer acetate (Copaxone), altered peptide ligands, mucosal administration of antigen and compounds that block co-stimulation pathways^{33,37,41}.



manner that regulates or suppresses inflammatory and even non-inflammatory processes that can cause tissue damage. Thus, one can redefine vaccination as 'the generation or induction of an immune response that is beneficial to the host in halting a pathological process', irrespective of whether that process is immune-mediated, autoimmune or even inflammatory.

Thus, vaccination involves not only the use of the immune system itself to correct or to alter abnormal immune responses that cause damage, but the immune system may be used to affect beneficially pathological processes that are neither autoimmune nor inflammatory. A striking example is represented by reports of the effectiveness of active immunization with A β peptide in adjuvant²⁹ and the passive administration of antibodies against A β ³⁰ to clear amyloid deposits and their surrounding glia and neuronal cytopathology from the brains of transgenic mouse models of AD.

It has also become clear that injury to the nervous system by non-immune mechanisms, such as stroke or trauma, may have a secondary stage associated with inflammation and that immune-based therapies can decrease CNS damage. For example, oral administration of MBP in a rat model of stroke decreases infarct size after middle cerebral artery occlusion and this is associated with increased expression of the anti-inflammatory cytokine TGF- β in the nervous system³¹, and nasal administration of myelin oligodendrocyte glycoprotein (MOG) has similar effects in a mouse model of stroke (D. Frenkel and H.W., unpublished results). In an extensive series of studies, Schwartz and co-workers³² have shown that T-cell autoimmunity against myelin antigens can be beneficial in animal models of central nervous system trauma caused by crush injury of the optic nerve or spinal cord contusion. Thus, an 'inflammatory response' directed against nervous system tissue also has the potential to have a protective or beneficial role.

Antigen-specific vaccination in MS

Antigen-specific modulation of the immune system is presumed to be the most specific and potentially least toxic way in which to manipulate the immune system in disease and represents the classic model of vaccination, that is, the induction of an antigen-specific beneficial immune response. For MS, a T_H1 -type cell-mediated disease, the strategy is to induce T_H2 or antigen-specific regulatory cells (Table 1 and Fig. 2b).

Numerous approaches using antigen-specific therapy have been successful in the murine EAE model and some of these have been tested in individuals with MS. The most successful so far has been the use of glatiramer acetate or copolymer 1, which is now an approved therapy for MS¹³. Glatiramer acetate is a random copolymer of four amino acids that was designed to mimic MBP and thus to induce EAE. It does not have encephalitogenic properties but instead works effectively in what seems to be an antigen-specific manner to suppress EAE by generating regulatory T cells. Although glatiramer acetate has several effects on the immune system, it seems principally to be acting as an altered peptide ligand that induces T_H2 - and T_H3 -type regulatory cells, which react in the CNS to suppress inflammation³³.

One of the major conceptual conundrums in designing antigen-specific vaccines for MS relates to the issue of which antigen to administer in MS. There is reactivity to several myelin antigens in MS^{34,35}, both because MS seems to be a syndrome rather than a single disease and because of epitope spreading, in which damage caused by a T cell specific for one myelin antigen induces reactivity to another myelin antigen³⁶. This conundrum seems to have been resolved by the phenomenon of bystander suppression, in which antigen-specific myelin-reactive regulatory cells are induced that secrete anti-inflammatory cytokines such as IL-10 and TGF- β (ref. 37). Such regulatory cells secrete anti-inflammatory cytokines when they encounter the autoantigen in the target tissue and thus suppress inflammation in the CNS caused by T cells of a different specificity. Thus, in the EAE model, one can suppress PLP-induced EAE by glatiramer acetate, by mucosal administration of MBP or by the use of altered peptide ligands of MBP, all of which induce anti-inflammatory regulatory T-cell responses (T_H2 , T_H3). Of note, in immune-deficient mice, T_H2 -type responses can induce a form of EAE³⁸.

But therapeutic vaccination is not without potential risks both in MS and in AD. In the early 1980s, Jonas Salk and colleagues attempted to treat individuals with MS by injecting large amounts of MBP subcutaneously to 'vaccinate' against putative harmful T-cell responses to MBP. They could induce both cellular and humoral (antibody) immune responses to MBP but obtained no consistent positive clinical effects and even some suggestion that the injections might have been harmful³⁹. To obviate harmful sensitization by injection of MBP, an analogous approach was undertaken using an altered

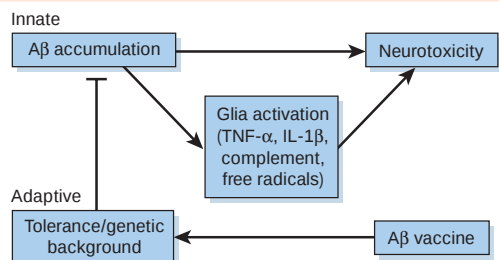


Figure 3 Inflammation and immune mechanisms in Alzheimer's disease.

Accumulation of A β leads to stimulation of the innate immune response, including activation of microglia and astrocytes, release of cytokines such as TNF- α and IL-1 β , complement activation and free-radical formation. This innate immune activation may contribute to neurotoxicity. An adaptive immune response induced by vaccination with A β generates antibodies against A β ; these antibodies decrease accumulation of A β in the brain through Fc-mediated clearance and also seem to draw A β from the brain into the cerebrospinal fluid and to the bloodstream. The adaptive immune response is under T-cell control and, depending on genetic background and T-cell immune tolerance, the effectiveness of vaccination to induce antibodies may be enhanced or decreased. In the context of A β vaccination and as part of the adaptive immune response, T_H1-type T cells directed against A β may be induced, which migrate to the nervous system and which may trigger an inflammatory response and a clinical picture of meningoencephalitis.

peptide ligand of MBP in which key amino acid sequences had been altered so that injection caused a T_H2 or T_H3 response as opposed to a T_H1 response. Results of a phase II trial showed that injections of large doses of the peptide led to a worsening of MS inflammation in some people, as measured by gadolinium-enhanced lesions on brain MRI, and an increased number of cells reactive to MBP⁴⁰. As part of a larger trial in individuals given a smaller dose, however, positive effects were observed on MRI and immune deviation towards T_H2-type responses was observed⁴¹.

As discussed below, an A β vaccine developed for use in AD has been found to cause adverse effects, which were most probably related to the induction of T_H1-type T-cell responses against A β . Of note, T-cell vaccination with myelin-reactive T cells to downregulate pathogenic T_H1 responses has been applied successfully to the EAE model and is being tested in individuals with MS^{28,42–44}, but it is not applicable to AD because there is no evidence of a pathogenic adaptive T-cell response in AD. DNA vaccination is another approach for treating CNS autoimmune diseases such as MS and has been used effectively in the EAE model by several investigators^{45–48}.

Alzheimer's disease

Alzheimer's disease is the most common form of age-related cognitive failure in humans. It is characterized neuropathologically by the progressive accumulation of the 42-residue A β peptide in limbic and association cortices, where some of it precipitates to form a range of amorphous and compacted extracellular plaques⁴⁹. These plaques, particularly the more compacted ones, are associated with dystrophic neurites (altered axons and dendrites), activated microglia and reactive astrocytes. Some of these dystrophic neurites contain intracellular bundles of abnormal paired helical filaments composed of insoluble, hyperphosphorylated forms of the microtubule-associated protein, tau. Paired helical filaments also accumulate in large cytoplasmic masses, called neurofibrillary tangles, in the cell bodies of innumerable limbic and neocortical neurons. The detection of neuritic (amyloid) plaques and neurofibrillary tangles in brain regions important for memory and other cognitive functions provides the basis for confirming a clinical diagnosis of AD after death.

Although it has become increasingly recognized that inflammation may be important in the neuropathological damage that occurs

in AD, unlike MS the inflammation in AD seems to arise from inside the CNS with little or no involvement of lymphocytes or monocytes beyond their normal surveillance of the brain^{50–52}. The inflammatory cytopathology (microgliosis, astrogliosis, complement activation, increased cytokine expression and acute phase protein response) is thought to represent a secondary response to the early accumulation of A β in the brain (Fig. 3). This innate immune response that occurs in the brain, which is presumably secondary to amyloid deposition, leads to the accumulation of inflammatory mediators such as TNF- α , IL-1, IL-6, free radicals and microglia activation.

To what degree this activation of microglia^{53,54} and other potential antigen-presenting CNS cells and secretors of cytokines is involved in the progressive neurodegenerative process is not yet clear, although it has been generally assumed to do more harm than good. Studies of transgenic mice that overexpress an AD-causing mutant form of human amyloid precursor protein (APP) and develop amyloid deposits have shown, however, that crossing such mice with mice overexpressing a natural inhibitor of complement C3 results in a worsening of A β plaque load and more neuronal loss⁵⁵. This result suggests that the inflammatory changes found in AD and mouse models thereof, including activation of the classical complement cascade, may represent a beneficial response, at least in part. Nonetheless, clinical studies suggest that conventional anti-inflammatory drugs such as those used in arthritis may delay or slow the progression of AD⁵⁰.

Despite the fact that only local innate inflammation occurs in AD, the theory and immune mechanisms of therapeutic vaccination discussed above with reference to MS have unexpectedly become relevant to AD, because the induction of specific adaptive immune responses has been shown to be of benefit in the animal model of AD (Fig. 3). It has been discovered that parenteral immunization of APP transgenic mice with synthetic A β in complete Freund's adjuvant can markedly decrease the number and density of A β deposits in the brain, with concomitant improvements in neuritic dystrophy and gliosis²⁹. Positive effects have also been found after repetitive mucosal (intranasal) administration of the peptide to transgenic mice⁵⁶. It seems that the induction of antibodies against A β has a primary role in the vaccine-mediated clearance of A β from the brain, because passive transfer of A β antibodies has shown similar beneficial neuropathological effects³⁰. Notably, a single parenteral administration of a monoclonal antibody against A β has been shown to produce rapid (within hours) benefits on certain behavioural measures of cognitive function in a mouse model, apparently by interfering with some diffusible, putatively synaptotoxic form of A β (for example, A β oligomers) without lowering the overall amount of A β deposits in the brain⁵⁷.

Two broad theories about the mechanisms by which A β antibodies work in mice have emerged. First, evidence of Fc-mediated uptake and clearance of A β antibody complexes by local activated microglia has been obtained³⁰. Second, evidence of a net movement of A β peptide out of the brain as a result of its binding and mobilization by A β antibodies, both peripherally (in the serum) and centrally (in the cerebrospinal fluid), has been provided⁵⁸. These two proposed mechanisms are not mutually exclusive, and there may be additional ways in which antibodies decrease A β -mediated synaptic and neuronal dysfunction. So far there is no clear evidence that T cells have either a protective or an injurious effect in AD or its mouse models, but this possibility needs further research. As discussed below, T-cell responses seem to have a role in the generation of meningoencephalitis after A β vaccination to induce antibodies.

Human trials of A β vaccination in AD

The finding that active vaccination with A β could profoundly reduce quantities of A β peptide in an animal model led to early clinical trials in which an A β _{1–42} synthetic peptide was administered parenterally with a previously tested adjuvant (QS21) to individuals with mild to moderate AD. Although a phase I safety study in few individuals did not detect significant side-effects, a subsequent phase II trial was

discontinued shortly after its initiation when roughly 5% of the treated participants developed what seemed to be an inflammatory reaction in the CNS (an aseptic meningoencephalitis). The occurrence of the meningoencephalitis was not correlated with either the presence or titres of antibodies against A β among the trial participants⁵⁹. The mechanism of this self-limited inflammatory reaction is unknown, but the appearance of the inflammation before the detection of A β antibodies in some of the recipients may suggest that a T-cell-mediated immune reaction to A β was responsible. Such cellular reactions were not detected in mice and other mammals exposed to the vaccine during preclinical safety and efficacy testing, although a recent report suggests autoimmune encephalomyelitis can be induced in mice vaccinated with A β peptide plus pertussis⁶⁰.

Efforts are underway to determine the basis for the adverse inflammatory reaction induced by A β _{1–42} and to attempt to model it in animals. No abnormal effects have been documented in APP transgenic mouse models to which A β antibodies have been administered, and such mice have shown robust clearing of brain A β deposits and even improvements in behavioural deficits^{30,57}. This is in contrast to the EAE model in which administration of antibody to MOG worsens the progression of EAE⁶¹. There is therefore an interest in conducting trials with a humanized monoclonal antibody to A β as the next step in the clinical evaluation of the immunotherapeutic approach to AD. It may also be possible to immunize with portions of A β to generate only antibodies that target N-terminal residues^{62,63}.

A β as an autoantigen

We have found recently that APP transgenic mice, which produce robust quantities of A β in the brain, have a form of immunological tolerance in which they show significantly lower T-cell responses when immunized with A β than do wild-type mice⁶³. This deficit can be overcome in part by providing T-cell help to the animal. Thus, the presence of abundant A β in the brain may not only cause local neuronal and glia damage but also hinder the generation of a therapeutic immune response, whether innate or induced.

Very recently, we have begun to extend such analyses to humans and, by using sensitive short-term cloning techniques, have found heightened *in vitro* reactivity of peripheral T-cells against A β in some elderly individuals and people with AD⁶⁴. Early studies did not find lymphocyte proliferation in response to APP peptides in individuals with AD⁶⁵. The likelihood of seeing this T-cell hyporeactivity in humans increased with age but was not observed in all individuals with AD or all aged normal individuals. Our results raise the possibility that endogenous T-cell reactivity in a host may relate to the progression of the cytopathological process of AD. In addition, such data suggest that it may be useful to test individuals for their intrinsic T-cell reactivity to A β before offering them any immunotherapeutic based on A β .

Beneficial versus deleterious T-cell responses

The issue of beneficial versus deleterious T-cell responses in vaccination models against CNS antigens is a concept that applies to approaches in both AD and MS. It has been shown that deleterious T-cell responses, presumably related to the induction of T_H1-type responses, can be induced in humans affected with either AD or MS. This does not mean that vaccination approaches in CNS diseases cannot be successful, as has been shown by the use of glatiramer acetate in MS; however, strategies that induce nonpathogenic T-cell responses must be utilized, for example, modified autoantigens, tolerogenic routes such as mucosal administration and non-T_H1-inducing adjuvants should be used, and careful attention should be paid to dosing. In addition, the genetic background of the host and the immune repertoire may also determine whether a detrimental T-cell response will occur after vaccination.

For example, we have found that SJL mice strains immunized with MOG peptide in complete Freund's adjuvant are susceptible to EAE, whereas B10S mice treated similarly are resistant⁶⁶. This does not seem

to relate to the generation of immune response against MOG, but to the type of immune response. In the SJL mouse there is infiltration of cells expressing γ -interferon in the brain and a predominantly T_H1 response, whereas in the B10S mice there is a T_H2 and T_H3 response that seems to prevent disease. Thus, the immune repertoire of the host before vaccination may determine the outcome of vaccination.

It seems that vaccination strategies both in AD and in MS will be dependent on skewing the immune response in such a way that it is not harmful to the host. In this regard, we have found in the APP mouse model of AD that nasal administration of A β induces antibody responses in association with an 'anti-inflammatory' cellular immune response involving IL-4, IL-10 and TGF- β ⁵⁶. These 'anti-inflammatory' responses may themselves help the pathologic process by suppressing inflammation and microglial activation, which are believed to contribute to the CNS dysfunction in AD^{50–52}. Furthermore, cells secreting TGF- β may themselves aid in the clearance of A β ⁵⁷. Such A β -reactive T cells would act in the CNS only at sites where A β is involved in the inflammatory process and thus would not be expected to interfere with normal physiology. □

doi:10.1038/nature01325

1. Prineas, J. W. in *Handbook of Multiple Sclerosis* (ed. Cook, S. D.) 223–255 (Marcel Dekker, Newark NJ, 1996).
2. Miller, D. H., Grossman, R. I., Reingold, S. C. & McFarland, H. F. The role of magnetic resonance techniques in understanding and managing multiple sclerosis. *Brain* **121**, 3–24 (1998).
3. Weiner, H. L. et al. Serial magnetic resonance imaging in multiple sclerosis: correlation with attacks, disability, and disease stage. *J. Neuroimmunol.* **104**, 164–173 (2000).
4. Cannella, B. & Raine, C. S. The adhesion molecule and cytokine profile of multiple sclerosis lesions. *Ann. Neurol.* **37**, 424–435 (1995).
5. Sorensen, T. L. et al. Expression of specific chemokines and chemokine receptors in the central nervous system of multiple sclerosis patients. *J. Clin. Invest.* **103**, 807–815 (1999).
6. Cuzner, M. L. et al. The expression of tissue-type plasminogen activator, matrix metalloproteases and endogenous inhibitors in the central nervous system in multiple sclerosis: comparison of stages in lesion evolution. *J. Neuropathol. Exp. Neurol.* **55**, 1194–1204 (1996).
7. Weiner, H. L. A 21 point unifying hypothesis on the etiology and treatment of multiple sclerosis. *Can. J. Neurolog. Sci.* **25**, 93–101 (1998).
8. Cook, S. D., Rohowsky-Kochan, C., Bansil, S. & Dowling, P. C. in *Handbook of Multiple Sclerosis* (ed. Cook, S. D.) 97–118 (Marcel Dekker, Newark, NJ, 1996).
9. Sibley, W. A., Bamford, C. R. & Clark, K. Clinical viral infections and multiple sclerosis. *Lancet* **1**, 1313–1315 (1985).
10. Edan, G. et al. Therapeutic effect of mitoxantrone combined with methylprednisolone in multiple sclerosis: a randomised multicenter study of active disease using MRI and clinical criteria. *J. Neurol. Neurosurg. Psychiatr.* **62**, 112–118 (1997).
11. Gobbi, M. I., Smith, M. E., Richert, N. D., Frank, J. A. & McFarland, H. F. Effect of open label pulse cyclophosphamide therapy on MRI measures of disease activity in five patients with refractory relapsing-remitting multiple sclerosis. *J. Neuroimmunol.* **99**, 142–149 (1999).
12. Weiner, H. L. & Cohen, J. A. Treatment of multiple sclerosis with cyclophosphamide: critical review of clinical and immunologic effects. *Mult. Scler.* **8**, 142–154 (2002).
13. Goodin, D. S. et al. Disease modifying therapies in multiple sclerosis: report of the Therapeutics and Technology Assessment Subcommittee of the American Academy of Neurology and the MS Council for Clinical Practice Guidelines. *Neurology* **58**, 169–178 (2002).
14. Mosmann, T. R. & Sad, S. The expanding universe of T-cell subsets: Th1, Th2, and more. *Immunol. Today* **17**, 138–146 (1996).
15. Maloy, K. J. & Powrie, F. Regulatory T cells in the control of immune pathology. *Nature Immunol.* **2**, 816–822 (2001).
16. Weiner, H. L. Induction and mechanism of action of transforming growth factor- β -secreting Th3 regulatory cells. *Immunol. Rev.* **182**, 207–214 (2001).
17. Groux, H. et al. A CD4⁺ T-cell subset inhibits antigen-specific T-cell responses and prevents colitis. *Nature* **389**, 737–742 (1997).
18. Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M. & Toda, M. Immunologic self-tolerance maintained by activated T cells expressing IL-2 receptor α -chains (CD25). Breakdown of a single mechanism of self-tolerance causes various autoimmune diseases. *J. Immunol.* **155**, 1151–1164 (1995).
19. Wucherpfennig, K. W. & Strominger, J. L. Molecular mimicry in T cell-mediated autoimmunity: viral peptides activate human T cell clones specific for myelin basic protein. *Cell* **80**, 695–705 (1995).
20. Panitch, H. S., Hirsch, R. L., Haley, A. S. & Johnson, K. P. Exacerbations of multiple sclerosis in patients treated with gamma interferon. *Lancet* **1**, 893–895 (1987).
21. Balashov, K. E., Smith, D. R., Khoury, S. J., Hafler, D. A. & Weiner, H. L. Increased IL-12 production in progressive multiple sclerosis: induction by activated CD4⁺ T-cells via CD40 ligand. *Proc. Natl Acad. Sci. USA* **94**, 599–603 (1997).
22. Windhagen, A. et al. Expression of costimulatory molecules B7-1 (CD80), B7-2 (CD86), and interleukin 12 cytokine in multiple sclerosis lesions. *J. Exp. Med.* **182**, 1985–1996 (1995).
23. Karni, A., Koldzic, D. N., Bharanidharan, P., Khoury, S. J. & Weiner, H. L. IL-18 is linked to raised IFN- γ in multiple sclerosis and is induced by activated CD4⁺ T cells via CD40–CD40 ligand interactions. *J. Neuroimmunol.* **125**, 134–140 (2002).
24. Balashov, K. E., Rottman, J. B., Weiner, H. L. & Hancock, W. W. CCR5⁺ and CXCR3⁺ T cells are increased in multiple sclerosis and their ligands MIP-1 α and IP-10 are expressed in demyelinating brain lesions. *Proc. Natl Acad. Sci. USA* **96**, 6873–6878 (1999).
25. Makhlof, K., Weiner, H. L. & Khoury, S. J. Increased percentage of IL-12⁺ monocytes in the blood correlates with the presence of active MRI lesions in MS. *J. Neuroimmunol.* **119**, 145–149 (2001).

26. Comabella, M. *et al.* Elevated interleukin-12 in progressive multiple sclerosis correlates with disease activity and is normalized by pulse cyclophosphamide therapy. *J. Clin. Invest.* **102**, 671–678 (1998).
27. Chabas, D. *et al.* The influence of the proinflammatory cytokine, osteopontin, on autoimmune demyelinating disease. *Science* **294**, 1731–1735 (2001).
28. Jansson, M., Panoutsakopoulou, V., Baker, J., Klein, L. & Cantor, H. Cutting edge: attenuated experimental autoimmune encephalomyelitis in η -1/osteopontin-deficient mice. *J. Immunol.* **168**, 2096–2099 (2002).
29. Schenk, D. *et al.* Immunization with amyloid- β attenuates Alzheimer-disease-like pathology in the PDAPP mouse. *Nature* **400**, 173–177 (1999).
30. Bard, F. *et al.* Peripherally administered antibodies against amyloid β -peptide enter the central nervous system and reduce pathology in a mouse model of Alzheimer disease. *Nature Med.* **6**, 916–919 (2000).
31. Becker, K. J. *et al.* Immunologic tolerance to myelin basic protein decreases stroke size after transient focal cerebral ischemia. *Proc. Natl Acad. Sci. USA* **94**, 10873–10878 (1997).
32. Moalem, G. *et al.* Autoimmune T cells protect neurons from secondary degeneration after central nervous system axotomy. *Nature Med.* **5**, 49–55 (1999).
33. Neuhaus, O., Farina, C., Wekerle, H. & Hohlfeld, R. Mechanisms of action of glatiramer acetate in multiple sclerosis. *Neurology* **56**, 702–708 (2001).
34. Olsson, T. *et al.* Autoreactive T lymphocytes in multiple sclerosis determined by antigen-induced secretion of interferon- γ . *J. Clin. Invest.* **86**, 981–985 (1990).
35. Zhang, J. *et al.* Increased frequency of IL-2 responsive T-cells specific for myelin basic protein and proteolipid protein in peripheral blood and cerebrospinal fluid of patients with multiple sclerosis. *J. Exp. Med.* **179**, 973–984 (1993).
36. Lehmann, P., Forsthuber, T., Miller, A. & Sercarz, E. Spreading of T-cell autoimmunity to cryptic determinants of an autoantigen. *Nature* **358**, 155–157 (1992).
37. Miller, A., Lider, O. & Weiner, H. L. Antigen-driven bystander suppression following oral administration of antigens. *J. Exp. Med.* **174**, 791–798 (1991).
38. Lafaille, J. J. *et al.* Myelin basic protein-specific T helper 2 (Th2) cells cause experimental autoimmune encephalomyelitis in immunodeficient hosts rather than protect them from the disease. *J. Exp. Med.* **186**, 307–312 (1997).
39. Romine, J. S. & Salk, J. in *Multiple Sclerosis: Pathology, Diagnosis and Management* (eds Hallpike, J. F., Adams, C. M. W. & Tourtelotte, W. W.) 621–630 (Williams & Wilkins, Baltimore, 1983).
40. Bielekova, B. *et al.* Encephalitogenic potential of the myelin basic protein peptide (amino acids 83–99) in multiple sclerosis: results of a phase II clinical trial with an altered peptide ligand. *Nature Med.* **6**, 1167–1175 (2000).
41. Kappos, L. *et al.* Induction of a non-encephalitogenic type 2 T helper-cell autoimmune response in multiple sclerosis after administration of an altered peptide ligand in a placebo-controlled, randomized phase II trial. The Altered Peptide Ligand in Relapsing MS Study Group. *Nature Med.* **6**, 1176–1182 (2000).
42. Kumar, V., Sercarz, E., Zhang, J. & Cohen, I. T-cell vaccination: from basics to the clinic. *Trends Immunol.* **22**, 539–540 (2001).
43. Vandenbark, A. A. *et al.* TCR peptide therapy in human autoimmune diseases. *Neurochem. Res.* **26**, 713–730 (2001).
44. Medaer, R., Stinissen, P., Truyen, L., Raus, J. & Zhang, J. Depletion of myelin-basic-protein autoreactive T cells by T-cell vaccination: pilot trial in multiple sclerosis. *Lancet* **346**, 807–808 (1995).
45. Weissert, R. *et al.* Protective DNA vaccination against organ-specific autoimmunity is highly specific and discriminates between single amino acid substitutions in the peptide autoantigen. *Proc. Natl Acad. Sci. USA* **97**, 1689–1694 (2000).
46. Wildbaum, G., Westermann, J., Maor, G. & Karin, N. A targeted DNA vaccine encoding fas ligand defines its dual role in the regulation of experimental autoimmune encephalomyelitis. *J. Clin. Invest.* **106**, 671–679 (2000).
47. Wildbaum, G., Netzer, N. & Karin, N. Plasmid DNA encoding IFN- γ -inducible protein 10 redirects antigen-specific T cell polarization and suppresses experimental autoimmune encephalomyelitis. *J. Immunol.* **168**, 5885–5892 (2002).
48. Garren, H. *et al.* Combination of gene delivery and DNA vaccination to protect from and reverse Th1 autoimmune disease via deviation to the Th2 pathway. *Immunity* **15**, 15–22 (2001).
49. Selkoe, D. J. Alzheimer's disease: genes, proteins and therapies. *Physiol. Rev.* **81**, 742–761 (2001).
50. Akiyama, H. *et al.* Inflammation and Alzheimer's disease. *Neurobiol. Aging* **21**, 383–421 (2000).
51. Cooper, N. R., Kalaria, R. N., McGeer, P. L. & Rogers, J. Key issues in Alzheimer's disease inflammation. *Neurobiol. Aging* **21**, 451–453 (2000).
52. McGeer, E. G. & McGeer, P. L. The importance of inflammatory mechanisms in Alzheimer disease. *Exp. Gerontol.* **33**, 371–378 (1998).
53. El Khoury, J., Hickman, S. E., Thomas, C. A., Loike, J. D. & Silverstein, S. C. Microglia, scavenger receptors, and the pathogenesis of Alzheimer's disease. *Neurobiol. Aging* **19**, S81–S84 (1998).
54. Meda, L. *et al.* Activation of microglial cells by β -amyloid protein and interferon- γ . *Nature* **374**, 647–650 (1995).
55. Wyss-Coray, T. *et al.* Prominent neurodegeneration and increased plaque formation in complement-inhibited Alzheimer's mice. *Proc. Natl Acad. Sci. USA* **99**, 10837–10842 (2002).
56. Weiner, H. L. *et al.* Nasal administration of amyloid- β peptide decreases cerebral amyloid burden in mouse model of Alzheimer's disease. *Ann. Neurol.* **48**, 567–579 (2000).
57. Dodart, J. C. *et al.* Immunization reverses memory deficits without reducing brain A β burden in Alzheimer's disease model. *Nature Neurosci.* **5**, 452–457 (2002).
58. DeMattos, R. B. *et al.* Peripheral anti-A β antibody alters CNS and plasma A β clearance and decreases brain A β burden in a mouse model of Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **98**, 8850–8855 (2001).
59. Schenk, D. in *Immunization against Alzheimer's Disease and Other Neurodegenerative Disorders* (eds Selkoe, D. & Christen, Y.) (Springer, Heidelberg, in the press).
60. Furlan, R. *et al.* Vaccination with amyloid- β peptide induces autoimmune encephalomyelitis in C57BL/6 mice. *Brain* (in the press).
61. Linington, C., Engelhardt, B., Kapocs, G. & Lassman, H. Induction of persistently demyelinated lesions in the rat following the repeated adoptive transfer of encephalitogenic T cells and demyelinating antibody. *J. Neuroimmunol.* **40**, 219–224 (1992).
62. McLaurin, J. *et al.* Therapeutically effective antibodies against amyloid- β peptide target amyloid- β residues 4–10 and inhibit cytotoxicity and fibrillogenesis. *Nature Med.* **11**, 1263–1269 (2002).
63. Monsonego, A., Maron, R., Zota, V., Selkoe, D. J. & Weiner, H. L. Immune hyporesponsiveness to amyloid β -peptide in amyloid precursor protein transgenic mice: implications for the pathogenesis and treatment of Alzheimer's disease. *Proc. Natl Acad. Sci. USA* **98**, 10273–10278 (2001).
64. Monsonego, A. *et al.* Immunogenic aspects of amyloid β -peptide: implications for pathogenesis and treatment of Alzheimer's disease. Abstr. no. 423 in Proc. VIII Int. Conf. on Alzheimer's Disease and Related Disorders, Stockholm, Sweden (2002).
65. Trieb, K., Ransmayr, G., Sgonc, R., Lassmann, H. & Grubeck-Lobenstein, B. APP peptides stimulate lymphocyte proliferation in normals, but not in patients with Alzheimer's disease. *Neurobiol. Aging* **17**, 541–547 (1996).
66. Maron, R. *et al.* Genetic susceptibility or resistance to autoimmune encephalomyelitis in MHC congenic mice is associated with differential production of pro- and anti-inflammatory cytokines. *Int. Immunol.* **11**, 1573–1580 (1999).
67. Wyss-Coray, T. *et al.* TGF- β 1 promotes microglial amyloid-beta clearance and reduces plaque burden in transgenic mice. *Nature Med.* **7**, 612–618 (2001).

The immunopathogenesis of sepsis

Jonathan Cohen

Brighton & Sussex Medical School, Westlain House, Falmer, Brighton BN1 9PH UK (e-mail: j.cohen@bsms.ac.uk)

Sepsis is a condition that results from a harmful or damaging host response to infection. Many of the components of the innate immune response that are normally concerned with host defences against infection can, under some circumstances, cause cell and tissue damage and hence multiple organ failure, the clinical hallmark of sepsis. Because of the high mortality of sepsis in the face of standard treatment, many efforts have been made to improve understanding of the dysregulation of the host response in sepsis. As a result, much has been learnt of the basic principles governing bacterial–host interactions, and new opportunities for therapeutic intervention have been revealed.

Sepsis describes a complex clinical syndrome that results from a harmful or damaging host response to infection. As a result of a concerted effort to understand the underlying pathogenetic mechanisms, there have been significant advances that have illuminated not just the process of sepsis, but also fundamental principles governing bacterial–host interactions. Unfortunately, attempts to translate these observations into improved clinical outcomes proved unsuccessful and led to considerable frustration. But in the past year, four major clinical trials that are based on somewhat different strategies have shown that it is possible to significantly reduce the mortality from sepsis and septic shock, and it is therefore timely to review these developments, both in basic science and its clinical applications.

Sepsis develops when the initial, appropriate host response to an infection becomes amplified, and then dysregulated. Clinically, the onset is often insidious: features may include fever, mental confusion, transient hypotension, diminished urine output or unexplained thrombocytopenia. If untreated, the patient may develop respiratory or renal failure, abnormalities of coagulation, and profound and unresponsive hypotension. A recent epidemiological study from North America found that the incidence was approximately 3.0 cases per 1,000 population, which translates into an annual burden of approximately 750,000 cases. The overall mortality is approximately 30%, rising to 40% in the elderly and is 50% or greater in patients with the more severe syndrome, septic shock¹. It is worth emphasizing that these figures represent mortality rates in patients admitted to hospital intensive care units and given antibiotics and the best available supportive care. The commonest sites of infection are the lungs, abdominal cavity, the urinary tract and primary infections of the blood stream. A microbiological diagnosis is made in about half the cases; Gram-negative bacteria account for about 60% of cases, Gram-positive for the remainder^{1,2}.

Microbial components that initiate injury

Determining the structural components of bacteria that are responsible for initiating the septic process has been important not only in understanding the underlying mechanisms, but also in identifying potential therapeutic targets. These bacterial motifs, which are recognized by the innate immune system, have been called pathogen-associated molecular patterns (PAMPs)³, although it might be more accurate to call them microorganism-associated molecular patterns as it is by no means clear how the host distinguishes between signals from pathogens rather than commensals.

In Gram-negative bacteria, lipopolysaccharide (LPS; known also as endotoxin) has a dominant role. The outer membrane of Gram-negative bacteria is constructed of a lipid bilayer, separated from the inner cytoplasmic membrane by peptidoglycan. The LPS molecule is embedded in the outer membrane and the lipid A portion of the molecule serves to anchor LPS in the bacterial cell wall.

Biophysical studies on the three-dimensional conformation adopted by different lipid A partial structures have revealed that, under physiological conditions, the most active forms assume the shape of a truncated cone, whereas inactive molecules prefer a lamellar structure and become progressively more cylindrical⁴. These conformational changes seem to correlate with the ability to activate host cell membranes.

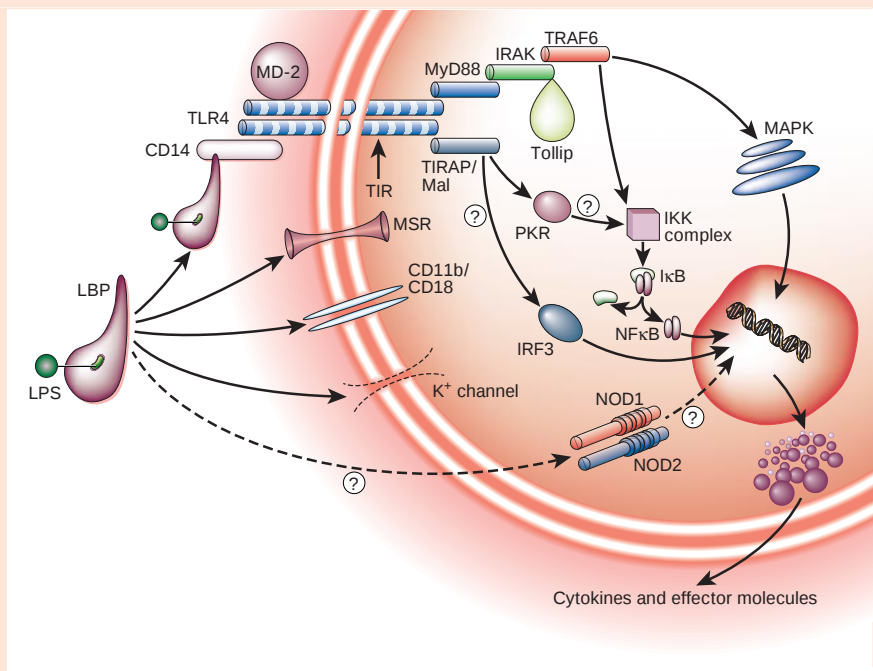
There is no endotoxin in Gram-positive bacteria, but their cell walls do contain peptidoglycan and lipoteichoic acid, and several investigators have identified structural components that account for their biological activity^{5,6}. Both peptidoglycan and lipoteichoic acid can bind to cell-surface receptors and are pro-inflammatory⁷, although they are much less active, on a weight-for-weight basis, than LPS. Their role in the pathogenesis of clinical sepsis remains uncertain because there are no convincing clinical data to show that they are present in the circulation at concentrations comparable to those used in the experimental setting.

However, an important feature of Gram-positive cells is the production of potent exotoxins, some of which are implicated in septic shock. The best known examples are the toxic shock syndromes caused by toxic shock syndrome toxin-1 (TSST-1)-producing strains of *Staphylococcus aureus* and the pyrogenic exotoxins from *Streptococcus pyogenes*. Toxic shock syndromes are among the most acute and most severe forms of septic shock. They frequently occur without warning in otherwise healthy individuals and the mortality can be as high as 50%. These Gram-positive exotoxins are of great interest because they exhibit the properties of superantigens, that is, they are able to bind promiscuously to major histocompatibility complex class II and a restricted repertoire of T-lymphocyte receptor (TCR) V β domains. In so doing they cause massive T-cell activation and release of pro-inflammatory lymphokines⁸, suggesting a plausible role for these toxins as a cause of the profound shock that is seen in patients with toxic shock.

Detailed structural analyses have been done for many bacterial superantigens, and the crystal structures of several staphylococcal and streptococcal toxins have been elucidated⁹. Interestingly, sequence variability in the amino-terminal domain dictates varying affinities for specific human leukocyte antigen (HLA) class II alleles; for instance, the strep-

Figure 1 Cell-surface recognition of lipopolysaccharide (LPS). The principal mechanism by which LPS is sensed is via an LPS-binding protein (LBP)–LPS complex and then signalling through the Toll-like receptor 4 (TLR4)–MD-2 complex. However, other cell surface molecules also sense LPS; these include the macrophage scavenger receptor (MSR), CD11b/CD18 and ion channels. Intracellular signalling depends on binding of the intracellular TLR domain, TIR (Toll/IL-1 receptor homology domain), to IRAK (IL-1 receptor-associated kinase), a process that is facilitated by two adapter proteins, MyD88 (myeloid differentiation protein 88) and TIRAP (TIR domain-containing adapter protein; also called MyD88-adapter-like protein or Mal), and inhibited by a third protein Tollip (Toll-interacting protein). Note that there is also an MyD88-independent pathway by which TIRAP/Mal signals through an RNA-dependent protein kinase (PKR) and interferon regulatory factor (IRF)-3. Recently it has been proposed that cells may also be able to respond to LPS by intracellular receptors called NOD proteins (for nucleotide-binding oligomerization domain). NOD1 (also called caspase-recruitment domain 4) was identified originally on the basis of structural homology to the apoptosis

regulator, Apaf-1. The NOD proteins have some similarities to the resistance (R) genes in plants that are involved in pathogen recognition; in common with TLRs and R genes, NODs have leucine-rich repeats. Expression of NOD1 and NOD2 confer responsiveness to Gram-negative LPS but not to lipoteichoic acid, which is found in Gram-positive bacteria. The mechanism by which NOD may recognize LPS in the cytosol is unknown.



tococcal superantigen SPEA (for streptococcal pyrogenic exotoxin A) shows significantly greater affinity for HLA-DQ than HLA-DR. These differences may in part explain the remarkable selectivity of the toxic shock syndromes: although staphylococcal and streptococcal strains bearing superantigen genes are widespread and indeed frequently cause infections, toxic shock syndromes are relatively uncommon.

Although experimental and epidemiological studies provide some support for the view that these superantigenic toxins are the cause of the toxic shock syndromes¹⁰, it is by no means clear that it is their superantigenicity *per se* that is responsible. For instance, despite many data that implicate the streptococcal toxin SPEA¹¹, this is in fact a relatively weak superantigen compared to the more recently described toxin streptococcal mitogenic exotoxin Z (SMEZ)¹². Yet in experimental models in which HLA-DQ transgenic mice are challenged with strains of *S. pyogenes* in which *smz* is disrupted, there is no effect on survival despite a profound reduction in pro-inflammatory activity¹³. These findings are important because there is considerable interest in devising therapeutic strategies that are targeted at Gram-positive infections and the toxic shock syndromes, and it is not clear whether these strategies should be aimed at the superantigenicity, or at other pro-inflammatory properties of the toxins.

There are also data that suggest that superantigenic toxins from Gram-positive bacteria induce hypersensitivity to LPS. The staphylococcal toxin TSST-1 enhances the susceptibility of rabbits to a lethal injection of LPS by a factor of approximately 50,000, and co-injection of LPS and TSST-1 induces tumour-necrosis factor- α (TNF- α) levels significantly higher than injection of similar doses of either toxin alone. Mice with severe combined immunodeficiency, lacking B and T lymphocytes, are resistant to this effect, but regain sensitivity when reconstituted with T-cells, and the mechanism seems to be dependent on enhanced production of interferon- γ (IFN- γ) from toxin-activated T cells¹⁴. This interaction between superantigens and LPS might in part explain the devastating nature of the toxic shock syndromes. It could also have therapeutic implications, as it might be advantageous to target LPS even if the infection is apparently caused exclusively by Gram-positive bacteria.

Several other bacterial components have been shown to have pro-inflammatory activity and to be able to induce shock in experimental systems. These include cell-wall structures such as flagellin¹⁵ and curli¹⁶, and unmethylated CpG sequences in naked bacterial DNA¹⁷. Receptors for some of these elements have been identified among the family of Toll-like proteins that are now known to be crucial in the cellular recognition of microbial structures¹⁸.

Host recognition of microbial components

The CD14–LBP complex

The inability to identify an ‘LPS receptor’ was for many years a barrier to understanding how Gram-negative bacteria initiated the septic response, but in a series of elegant studies it was shown that activation of host cells was dependent on the presence of LPS-binding protein (LBP) and the opsonic receptor CD14 (ref. 19). Although CD14 was originally identified as the essential co-receptor that mediated LPS activation of monocytes, subsequent work has shown that it also participates in the activation by Gram-positive cell-wall components such as peptidoglycan²⁰, mediates macrophage apoptosis²¹, and is important in shuttling LPS between serum proteins that have the capacity to bind LPS, such as LBP and serum lipoproteins²². Membrane bound CD14 (mCD14) is a glycosylphosphatidylinositol-linked molecule anchored in the cell surface, but it is also found in the circulation as soluble CD14 (sCD14). Many cells that are constitutively CD14 negative, such as dendritic cells, fibroblasts, smooth muscle cells and vascular endothelium, are still able to respond to LPS by interacting with sCD14. sCD14 is found in the serum of healthy individuals but levels rise in sepsis²³, and antibody to CD14 protects primates from lethal endotoxin shock²⁴.

Toll-like receptors

Although the discovery of CD14 represented a significant step forward in understanding host responses to LPS, the fact that mCD14 had no intracellular tail meant that it remained unclear how ligation of the LPS–LBP complex led to cellular activation. This uncertainty was resolved by the discovery of the family of Toll-like receptors

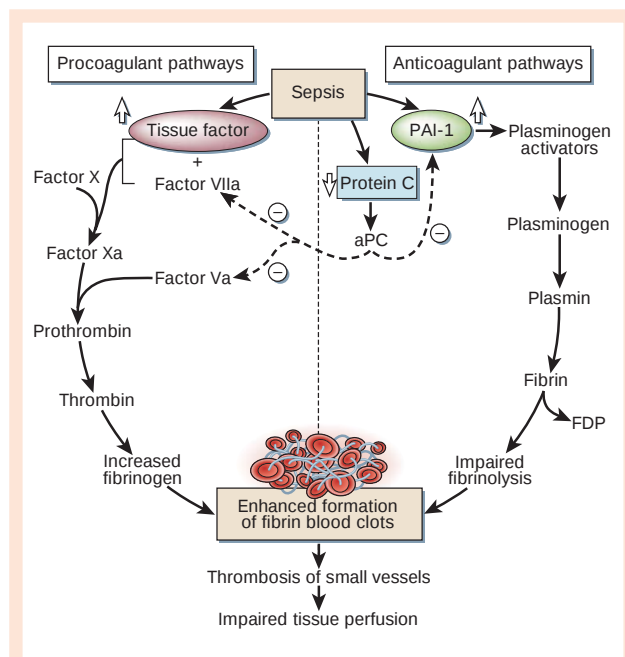


Figure 2 Sepsis disturbs the normal homeostatic balance between procoagulant and anticoagulant mechanisms. Tissue factor expression is enhanced leading to increased production of prothrombin that is converted to thrombin, and that in turn generates fibrin from fibrinogen. Simultaneously, levels of the plasminogen-activator inhibitor-1 (PAI-1) are increased, resulting in impaired production of plasmin and thus failure of normal fibrinolytic mechanisms by which fibrin is converted to degradation products (FDP). Sepsis also causes a fall in the levels of the natural anticoagulant protein C (and also antithrombin and the tissue factor pathway inhibitor, TFPI, not shown). The activated form of protein C, aPC, dissociates from the endothelial protein C receptor to inactivate factors Va and VIIa and inhibit PAI-1 activity; hence reduced levels of protein C result in further procoagulant effect. The net result is enhanced formation of fibrin clots in the microvasculature, leading to impaired tissue oxygenation and cell damage.

(TLRs)^{25,26}. Over a remarkably short period of time, studies of innate immunity in *Drosophila* revealed the existence of a proteolytic cascade that yielded ligands for cellular receptors that could distinguish bacterial from fungal infection. It was shown subsequently that there were striking similarities between this system and the interleukin (IL)-1 signalling system in mammals. This in turn led to the identification of human TLRs²⁷ and the discovery that a TLR was the long-sought co-receptor for LPS²⁸.

A family of (currently) ten TLRs has been identified with a wide range of ligand specificity including bacterial, fungal and yeast proteins^{25,29}. Thus, TLR4 is the LPS receptor, TLR2 is predominantly responsible for recognizing Gram-positive cell-wall structures³⁰, TLR5 is the receptor for flagellin³¹ and TLR9 recognizes CpG elements in bacterial DNA¹⁸. An additional cell-surface molecule, MD-2, has been identified that is required for activation of TLR4 (ref. 32). MD-2 knockout mice do not respond to LPS and survive endotoxic shock. The role of MD-2 seems to be that of positioning TLR4 correctly on the cell surface, as in MD-2^{-/-} embryonic fibroblasts TLR4 remained within the Golgi and failed to appear on the cell surface³³.

The notion of a 'monogamous' association between one particular TLR and its microbial ligand, as in the case of LPS and TLR4, is in reality an oversimplification. For instance, TLR2 can be activated by cell-wall components of both yeast and mycobacteria. Further complexity is introduced into the system by the fact that TLRs seem to be able to combine to form a repertoire capable of distinguishing closely related ligands³⁴, and there is at least preliminary evidence that polymorphisms in Toll-family proteins might provide part of the

explanation for the enormous variability in individual responses to what seem to be similar infective challenges^{35,36}.

Signalling pathways activated by TLRs have been dissected in great detail, and show a remarkable degree of homology with the Toll activation pathway in *Drosophila*³⁷. TLRs have an intracellular domain that is homologous with the IL-1 receptor and the IL-18 receptor. Adapter proteins facilitate binding to IL-1 receptor-associated kinase, which in turn induces TNF receptor-associated factor-6, leading to nuclear translocation of nuclear factor- κ B (NF- κ B) and ultimately to activation of cytokine gene promoters (Fig. 1). Although this model is based on LPS signalling of TLR4, a similar — although not identical — process is involved in the activation of TLR2 by Gram-positive bacteria.

Other host signal molecules that respond to bacteria

A further layer of complexity has been provided by the discovery that there are several additional pathways by which cells recognize microbial components. Peptidoglycan-recognition proteins (PGRPs) were identified in moths and subsequently a family of PGRP genes was found in *Drosophila*³⁸ and in humans³⁹. Different PGRPs can distinguish between Gram-positive⁴⁰ and Gram-negative bacteria^{41,42}. In *Drosophila*, they seem to act by regulating activation of Relish, a member of the NF- κ B family⁴³, although the precise mechanism by which they are sensed at the cell surface remains unknown.

The triggering receptor expressed on myeloid cells (TREM-1) and the myeloid DAP12-associating lectin (MDL-1) are two recently identified receptors involved in monocyte activation and inflammatory response. TREM-1 is upregulated in the presence of various microorganisms⁴⁴, although the ligand for TREM-1 is unknown. When mononuclear cells are exposed to a combination of LPS and an antibody to TREM-1, there is a synergistic effect and enhanced production of pro-inflammatory cytokines. But if a fusion protein of TREM-1 and the Fc portion of IgG is used to compete with cell-bound receptor, LPS-induced cytokine production is downregulated and mice can be protected from death up to 4 hours after a lethal injection of LPS⁴⁵. This is a therapeutic effect that will have obvious implications if it can be reproduced in clinical studies.

Finally, there is the recent description of the monocyte intracellular proteins NOD1 and NOD2 (for nucleotide-binding oligomerization domain), which seem to have the ability to bind and to confer responsiveness to LPS⁴⁶, suggesting that this might be yet another way cells respond to the presence of bacteria⁴⁷. Genotypic variations in NOD2 are associated with distinct clinical phenotypes of Crohn's disease⁴⁸, prompting speculation that other NOD genotypes might be associated with phenotypic variations in LPS responsiveness.

Signal amplification

Following the initial host-microbial interaction there is widespread activation of the innate immune response, the purpose of which is to coordinate a defensive response involving both humoral and cellular components. Mononuclear cells play a key role, releasing the classic pro-inflammatory cytokines IL-1, IL-6 and TNF- α , but in addition an array of other cytokines including IL-12, IL-15 and IL-18, and a host of other small molecules (Table 1).

TNF- α and IL-1 are the prototypic inflammatory cytokines that mediate many of the immunopathological features of LPS-induced shock⁴⁹. They are released during the first 30–90 minutes after exposure to LPS and in turn activate a second level of inflammatory cascades including cytokines, lipid mediators and reactive oxygen species, as well as upregulating cell adhesion molecules that result in the initiation of inflammatory cell migration into tissues. The fact that anti-TNF or anti-IL-1 strategies failed to prevent death in septic patients is probably related more to the difficulty of designing clinical trials in these patients, rather than an intrinsic flaw in the scientific rationale⁵⁰. One practical problem is that patients often come to medical attention relatively late in the disease, and blocking these early cytokines may simply be too late. High mobility group B1

(HMGB1) has recently been identified as a cytokine-like product of macrophages that appears much later after LPS stimulation and may represent a more tractable target for intervention⁵¹.

HMGB1 is a non-histone chromosomal protein that is abundantly distributed and exists in nuclear, cytoplasmic and membrane-bound forms. It participates in stabilizing nucleosomes, facilitates gene transcription and modulates the activity of steroid hormone receptors. When mice were injected with LPS, HMGB1 serum concentrations rose after a delay of about 24 hours, long after the initial peak of IL-1 and TNF- α had declined. Importantly, mice could be rescued from LPS-induced shock by administering an antibody to HMGB1, even when this was provided up to 2 hours after the lethal injection⁵². Subsequently it was shown that patients with sepsis have elevated serum levels of HMGB1, and that higher levels are associated with an increased mortality, suggesting that clinical intervention by blocking or neutralizing HMGB1 might be a viable option.

Another macrophage-derived cytokine that has been identified as a potential therapeutic target in sepsis is macrophage migration inhibitory factor (MIF). Mice with a targeted disruption of the MIF gene are resistant to LPS-induced shock⁵³ and antibody to MIF is fully protective, even in the more demanding caecal ligation and puncture model that resembles clinical peritonitis⁵⁴. MIF also seems to mediate shock caused by Gram-positive bacteria, such as the toxic shock syndrome associated with *S. aureus*⁵⁵, suggesting that anti-MIF strategies might have broad application in septic patients. MIF has a curious relationship with glucocorticoids, which are normally thought of as being anti-inflammatory, as low doses of glucocorticoids paradoxically induce macrophage MIF. Once released, MIF then acts as a pro-inflammatory agent, over-riding the ability of glucocorticoids to prevent shock in animal models of sepsis⁵⁶. How this complex relationship manifests in a clinical setting is of particular interest in the light of the recent studies demonstrating a protective effect of low-dose steroids in patients with severe sepsis.

These pro-inflammatory cytokines are important because they in turn are responsible for orchestrating a complex network of secondary responses (for a review, see ref. 49). A good example of this is provided by IL-18, a cytokine that induces production of interferon- γ (IFN- γ). In human mononuclear cells, IFN- γ upregulates surface expression of TLR4, MD-2 and MyD88, and counteracts the LPS-induced downregulation of TLR4 (ref. 57). It has long been known that IFN- γ sensitizes human mononuclear cells to the effects of LPS, and these new findings suggest strongly that this effect is probably mediated through upregulation (or at least, prevention of downregulation) of TLR4.

The coagulation cascade

Cytokines are also important in inducing a procoagulant effect in sepsis. Disorders of coagulation are common in sepsis, and 30–50% of patients have the more severe clinical form, disseminated intravascular coagulation⁵⁸. Coagulation pathways are initiated by LPS and other microbial components, inducing expression of tissue factor on mononuclear and endothelial cells. Tissue factor in turn activates a series of proteolytic cascades, which result in the conversion of prothrombin to thrombin, which in turn generates fibrin from fibrinogen. Simultaneously, normal regulatory fibrinolytic mechanisms (fibrin breakdown by plasmin) are impaired because of high plasma levels of plasminogen-activator inhibitor type-1 (PAI-1) that prevent the generation of plasmin from the precursor plasminogen. The net result is enhanced production and reduced removal of fibrin leading to the deposition of fibrin clots in small blood vessels, inadequate tissue perfusion and organ failure (Fig. 2).

Pro-inflammatory cytokines, in particular IL-1 and IL-6, are powerful inducers of coagulation, and conversely, IL-10 regulates coagulation by inhibiting the expression of tissue factor on monocytes (for a review, see ref. 59). An additional cause of the procoagulant state in sepsis is the downregulation of three naturally occurring anticoagulant proteins — antithrombin, protein C and tissue factor

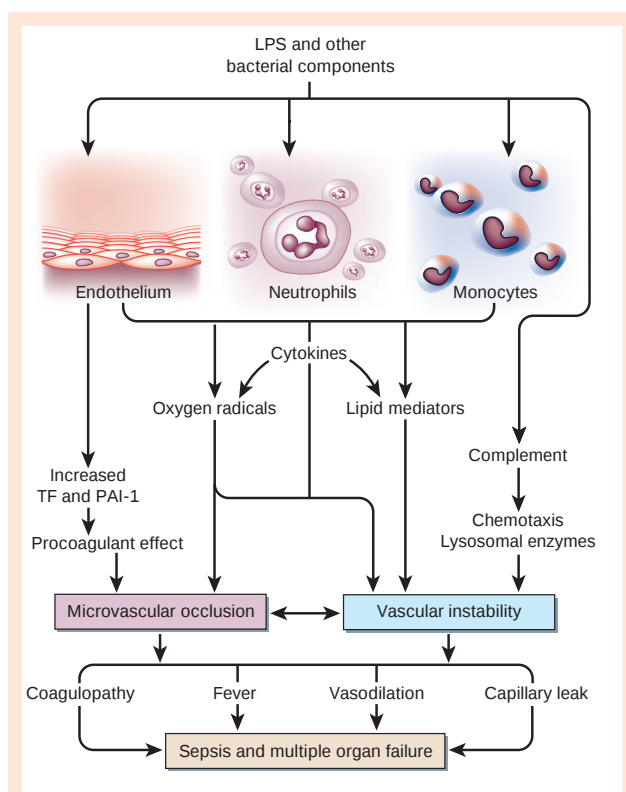


Figure 3 Pathogenetic networks in shock. Lipopolysaccharide (LPS) and other microbial components simultaneously activate multiple parallel cascades that contribute to the pathophysiology of adult respiratory distress syndrome (ARDS) and shock. The combination of poor myocardial contractility, impaired peripheral vascular tone and microvascular occlusion leads to tissue hypoperfusion and inadequate oxygenation, and thus to organ failure.

pathway inhibitor. These natural anticoagulants are of particular interest because in addition to their effect on thrombin generation, they also have anti-inflammatory properties, including effects on release of monocyte-derived TNF- α by inhibiting activation of the transcription factors NF- κ B and activator protein (AP)-1 (ref. 60).

Particular attention has focused on Protein C, which is converted to the activated form (aPC) when thrombin complexes with thrombomodulin, an endothelial transmembrane glycoprotein. Once aPC is formed it dissociates from an endothelial protein C receptor (EPCR) before binding protein S, resulting in inactivation of factors Va and VIIIa and thus blockade of the coagulation cascade. It has been shown recently that aPC uses EPCR as a co-receptor for cleavage of protease-activated receptor 1 (PAR1). Gene profiling showed that PAR1 signalling could account for the activation of aPC-induced protective genes, including the immunomodulatory monocyte chemoattractant protein-1 (MCP-1), suggesting a role for PAR-1 activation in protection from sepsis⁶¹. In septic patients, aPC levels are reduced and expression of endothelial thrombomodulin and EPCR are impaired⁶², providing some support for the notion that replacement of aPC might have therapeutic value.

The counter-inflammatory response — modifier or mediator?

The profound pro-inflammatory response that occurs in sepsis is balanced by an array of counter-regulatory molecules that attempt to restore immunological equilibrium. In this sense, the counter-inflammatory response is seen as a 'modifier' — both appropriate and beneficial. Counter-inflammatory cytokines include antagonists such as the soluble TNF receptors and IL-1 receptor antagonist, decoy receptors such as IL-1 receptor type II, inactivators of the complement

Table 1 Macrophage products implicated in the pathogenesis of sepsis

Mediators	Typical effects
Cytokines	
IL-1, IL-6, IL-12, IL-15, IL-18, TNF- α , MIF, HMGB1, IL-10	Activate neutrophils, lymphocytes and vascular endothelium; upregulate cellular adhesion molecules; induce prostaglandins, nitric oxide synthase and acute-phase proteins; induce fever. Note that IL-10 is predominantly a negative regulator of these effects
Chemokines	
IL-8, MIP-1 α , MIP-1 β , MCP-1, MCP-3	Mobilize and activate inflammatory cells, especially neutrophils; activate macrophages
Lipid mediators	
Platelet-activating factor, prostaglandins, leukotrienes, thromboxane, tissue factor	Activate vascular endothelium; regulate vascular tone; activate extrinsic coagulation cascade
Oxygen radicals	
Superoxide and hydroxyl radicals, nitric oxide	Antimicrobial properties; regulation of vascular tone

Upon activation by microbial components, macrophages release a diverse range of products; the table lists the principal among them that are implicated in the pathogenesis of sepsis. Most of these products have multiple targets and cause effects through several parallel mechanisms. For instance, nitric oxide has direct antimicrobial properties, can modify vascular tone (and hence tissue oxygenation), and also causes macrophage apoptosis. Furthermore, many macrophage products are involved in the regulation of each other. Thus, TNF- α upregulates tissue factor and nitric oxide synthase, IL-18 induces IFN- γ , which in turn further activates macrophages, while IL-10 is a global suppressor of macrophage function. These highly complex and tightly regulated networks make it difficult to predict the consequences of blocking or inhibiting just one pathway.

cascade and the anti-inflammatory cytokines, of which the prototype is IL-10. In concert with this, the host response to injury includes profound changes in metabolic activity (increased cortisol production and release of catecholamines), induction of acute-phase proteins, and endothelial activation with upregulation of adhesion molecules and release of prostanoids and platelet-activating factor (PAF).

Another facet of downregulation of immunity that occurs in sepsis is the development of lymphocyte apoptosis. Extensive lymphocyte apoptosis is seen in animal models of sepsis and is also present in septic patients, although interestingly, much less so in critically ill non-septic controls⁶³. Septic patients are usually lymphopenic, and subset analysis of autopsy tissue samples has shown that there is selective depletion of B and CD4⁺ lymphocytes⁶⁴. This process and its functional consequences are viewed as part of a more general state of immunosuppression, characterized by T-cell hyporesponsiveness and anergy, which occurs to some extent in most septic patients, and which is seen as a counter-balancing response (and sometimes, over-response) to the initial pro-inflammatory state.

It is because of this over-response that some investigators view the counter-inflammatory response as the cause of an inadequate host defence against infection and hence a potential 'mediator' of sepsis and progressive organ failure. Several have pursued the notion that reversal of this immunosuppressive state might be of therapeutic value. For instance, mice transfected with the human gene *bcl-2* that overexpress

the anti-apoptotic protein Bcl-2 are protected from death after caecal ligation and puncture⁶⁵, and patients that received IFN- γ in a small non-randomized clinical study showed upregulation of HLA-DR on their monocytes and a better-than-anticipated survival⁶⁶.

Role of genetic susceptibility in the pathogenesis of sepsis

Among this vast array of host molecules that orchestrate the response to sepsis there are many examples of genetic variability that influence physiological activity. For example, there has been great interest in exploring the possibility that a polymorphism in the TNF promoter that results in significantly higher TNF levels might be associated with a worse outcome from sepsis. Several of these associations have been studied⁶⁷ (Table 2) and at least in some cases the evidence seems convincing. Of particular interest was the recent report that mutations in TLR4 are associated with an increased susceptibility to Gram-negative sepsis³⁶.

Mechanisms of organ failure

The ultimate cause of death in patients with sepsis is multiple organ failure. Typically, patients will first develop a single organ failure — for instance, respiratory failure requiring mechanical ventilation — and then if the disease remains unchecked, will progressively develop failure of other organ systems. There is a close relationship between the severity of organ dysfunction on admission to an intensive care unit and the probability of survival, and between the numbers of organs failing and the risk of death. If four or five organs fail the mortality is greater than 90%, irrespective of treatment.

The pathogenesis of organ dysfunction is multifactorial and incompletely understood. Tissue hypoperfusion and hypoxia are dominant factors (Fig. 3). The mechanisms involve widespread fibrin deposition causing microvascular occlusion, the development of tissue exudates further compromising adequate oxygenation, and disorders of microvascular homeostasis resulting from the elaboration of vasoactive substances such as PAF, histamine and prostanoids. Cellular infiltrates, particularly neutrophils, damage tissue directly by releasing lysosomal enzymes and superoxide-derived free radicals. TNF- α and other cytokines increase the expression of the inducible nitric oxide synthase and increased production of nitric oxide causes further vascular instability and may also contribute to the direct myocardial depression that occurs in sepsis⁶⁸.

The tissue hypoxia that develops in sepsis is reflected in the oxygen debt — that is, the difference between oxygen delivery and oxygen requirements. The extent of the oxygen debt is related to the outcome from sepsis, and strategies designed to optimize oxygen delivery to the tissues can improve survival. In addition to hypoxia, cells may be dysoxic — that is, unable to properly utilize available oxygen. Recent data suggest that this may be another consequence of excess nitric oxide production, because skeletal muscle biopsies from septic

Table 2 Summary of genetic polymorphisms that have been explored in patients with sepsis and meningococcal disease

Host factor	Genetic association	Clinical effect
Tumour-necrosis factor- α (TNF- α)	TNF2 allele G308A mutation in TNF promoter	Increased serum levels of TNF associated with non-survival in sepsis
Lymphotoxin- α (TNF- β)	TNFB2 allele	Increased levels of lymphotoxin associated with sepsis in trauma patients
Interleukin-1 receptor antagonist (IL-1Ra)	ILRN*2 allele	Associated with increased levels IL-1 β and more severe sepsis. Homozygotes for ILRN*2 and TNFB2 have a higher mortality from sepsis
Factor V Leiden	FV ^L	Possibly a mild effect on the severity of meningococcal disease
Tissue plasminogen activator (t-PA)	Insertion/deletion polymorphism	No effect in meningococcal disease
Plasminogen-activator inhibitor 1 (PAI-1)	4G/5G insertion/deletion polymorphism	4G/4G haplotype associated with increased levels of PAI-1 and increased severity of meningococcal disease
FC γ RIIIa (CD32)	R131 allotype	More severe outcome from meningococcal disease
Lipopolysaccharide-binding protein (LBP) and bactericidal/permeability-increasing protein (BPI)	Various	No statistically significant effect in sepsis
CD14	C(-159)T promoter polymorphism	Increased susceptibility to septic shock and death
Toll-like receptor 4 (TLR4)	Asp299Gly allele	Significant association with septic shock, and in particular Gram-negative sepsis

It should be noted that not all these associations have been confirmed independently, and the significance of the findings is influenced by the precise patient population studied and the methods used. Other host factors (for instance, mannan binding protein) are known to be important in the response to specific infections, but have not yet been studied specifically in sepsis; yet others (for example, nitric oxide synthase) have been suggested to be important, but clinical data are lacking.

patients show evidence of impaired mitochondrial respiration, which is inhibited by nitric oxide⁶⁹.

Therapeutic approaches

Despite the extraordinary developments in understanding the immunopathology of sepsis, therapeutic advances have been painfully slow. However, in the past 12 months several clinical trials have finally shown that it is possible to reduce mortality in patients with sepsis. Interestingly, each has been based on a different aspect of the pathology described above.

The critical importance of tissue oxygenation was addressed by a study in which patients in the earliest stages of sepsis were treated by aggressive management with fluids, blood transfusion and inotropic agents to optimize haemodynamic function⁷⁰. An alternative approach was taken by van den Bergh and co-workers, who studied the effect of rigorous control of blood glucose levels. Hyperglycaemia and insulin resistance are common in critically ill patients, even if they have not previously had diabetes, and these authors showed that intensive insulin therapy could substantially reduce mortality⁷¹. The mechanism of this striking effect is not absolutely clear, although it is of interest that the greatest reduction in mortality involved deaths due to multiple-organ failure in patients with a proven septic focus, perhaps suggesting that it was related to the better control of the initial infective process.

The third study to show significant benefit was a trial of low-dose corticosteroids. Earlier trials using very high doses of steroids, based on the premise that sepsis represented an uncontrolled inflammatory response, had failed to show any survival benefit. But Annane and colleagues had noted that patients in the advanced stages of septic shock had relative adrenal insufficiency and reasoned that low-dose replacement steroids might be beneficial. In a phase III trial in highly selected patients, they found that low doses of hydrocortisone and fludrocortisone did indeed reduce the mortality substantially⁷². But doubts remain about the precise mechanism of this effect and it is possible that the benefit derives, at least in part, from the immunosuppressive effects of the hydrocortisone.

Disordered coagulation is important in sepsis, and the fourth study examined the effects on the course of sepsis of replacing aPC. A randomized controlled trial of aPC therapy resulted in a significant survival benefit in treated patients⁷³, although interestingly the effect was equally marked in patients whose protein C levels were not depressed. This suggests that the therapeutic effect was attributable in part to the immunosuppressive effects of protein C⁶⁰, and not just to its anticoagulant properties.

Improved understanding of the immunopathology of sepsis has facilitated many other approaches, and several additional strategies are at various stages of development. These include therapies aimed at bacterial targets, for example novel anti-endotoxin molecules such as bactericidal/permeability-increasing protein, or modified lipoproteins, both of which absorb and neutralize LPS, as well as very recent reports that oxidized phospholipids can interfere with binding of LPS to LBP⁷⁴, and strategies aimed at Gram-positive toxins, including competitive antagonists of superantigen-binding sites. There are also investigations aimed at host molecules, such as PAF-receptor antagonists, and a variety of targets in the coagulation cascade. For the clinical investigator, the challenge will be the design of studies with sufficient power to determine the potential value of these new therapies. As fatality rates drop with the use of low-dose steroids, aPC and the other measures described above, it will become increasingly difficult to carry out studies on a heterogeneous population of patients with sepsis. The way forward is likely to lie in identifying clinically relevant endpoints other than death (for instance, reduced incidence of organ failure), and/or identifying more homogeneous subgroups of patients in whom to study specifically targeted therapeutic interventions. □

doi:10.1038/nature01326

1. Angus, D. C. *et al.* Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Crit. Care Med.* **29**, 1303–1310 (2001).

2. Alberti, C. *et al.* Epidemiology of sepsis and infection in ICU patients from an international multicentre cohort study. *Intensive Care Med.* **28**, 108–121 (2002).
3. Janeway, C. A. Jr & Medzhitov, R. Introduction: the role of innate immunity in the adaptive immune response. *Semin. Immunol.* **10**, 349–350 (1998).
4. Seydel, U., Oikawa, M., Fukase, K., Kusumoto, S. & Brandenburg, K. Intrinsic conformation of lipid A is responsible for agonistic and antagonistic activity. *Eur. J. Biochem.* **267**, 3032–3039 (2000).
5. Majcherczyk, P. A. *et al.* Digestion of *Streptococcus pneumoniae* cell walls with its major peptidoglycan hydrolase releases branched stem peptides carrying proinflammatory activity. *J. Biol. Chem.* **274**, 12537–12543 (1999).
6. Morath, S., Geyer, A. & Hartung, T. Structure–function relationship of cytokine induction by lipoteichoic acid from *Staphylococcus aureus*. *J. Exp. Med.* **193**, 393–397 (2001).
7. Wang, J. E. *et al.* Peptidoglycan and lipoteichoic acid from *Staphylococcus aureus* induce tumor necrosis factor alpha, interleukin 6 (IL-6), and IL-10 production in both T cells and monocytes in a human whole blood model. *Infect. Immun.* **68**, 3965–3970 (2000).
8. Lavoie, P. M., Thibodeau, J., Erard, F. & Sekaly, R. P. Understanding the mechanism of action of bacterial superantigens from a decade of research. *Immunol. Rev.* **168**, 257–269 (1999).
9. Papageorgiou, A. C. & Acharya, K. R. Microbial superantigens: from structure to function. *Trends Microbiol.* **8**, 369–375 (2000).
10. Korb, M. Bacterial pyrogenic exotoxins as superantigens. *Clin. Microbiol. Rev.* **8**, 411–426 (1995).
11. Norrby-Teglund, A. *et al.* Evidence for superantigen involvement in severe group A streptococcal tissue infections. *J. Infect. Dis.* **184**, 853–860 (2001).
12. Muller-Alouf, H. *et al.* Pyrogenicity and cytokine-inducing properties of *Streptococcus pyogenes* superantigens: comparative study of streptococcal mitogenic exotoxin Z and pyrogenic exotoxin A. *Infect. Immun.* **69**, 4141–4145 (2001).
13. Unnikrishnan, M. *et al.* The bacterial superantigen streptococcal mitogenic exotoxin Z is the major immunoinactive agent of *Streptococcus pyogenes*. *J. Immunol.* **169**, 2561–2569 (2002).
14. Dinges, M. M. & Schlievert, P. M. Role of T cells and gamma interferon during induction of hypersensitivity to lipopolysaccharide by toxic shock syndrome toxin 1 in mice. *Infect. Immun.* **69**, 1256–1264 (2001).
15. Eaves-Pyles, T. *et al.* Flagellin, a novel mediator of salmonella-induced epithelial activation and systemic inflammation: IkB degradation, induction of nitric oxide synthase, induction of proinflammatory mediators, and cardiovascular dysfunction. *J. Immunol.* **166**, 1248–1260 (2001).
16. Herwaldt, H. *et al.* Activation of the contact-phase system on bacterial surfaces—a clue to serious complications in infectious diseases. *Nature Med.* **4**, 298–302 (2001).
17. Sparwasser, T. *et al.* Bacterial DNA causes septic shock. *Nature* **386**, 336–337 (1997).
18. Bauer, S. *et al.* Human TLR9 confers responsiveness to bacterial DNA via species-specific CpG motif recognition. *Proc. Natl Acad. Sci. USA* **98**, 9237–9242 (2001).
19. Wright, S. D., Ramos, R. A., Tobias, P. S., Ulevitch, R. J. & Mathison, J. C. CD14, a receptor for complexes of lipopolysaccharide (LPS) and LPS binding protein. *Science* **249**, 1431–1433 (1990).
20. Pugin, J. *et al.* CD14 is a pattern recognition receptor. *Immunology* **1**, 509–516 (1994).
21. Devitt, A. *et al.* Human CD14 mediates recognition and phagocytosis of apoptotic cells. *Nature* **392**, 505–509 (1998).
22. Yu, B., Hallman, E. & Wright, S. D. Lipopolysaccharide binding protein and soluble CD14 catalyze exchange of phospholipids. *J. Clin. Invest.* **99**, 315–324 (1997).
23. Landmann, R. *et al.* Increased circulating soluble CD14 is associated with high mortality in Gram-negative septic shock. *J. Infect. Dis.* **171**, 639–644 (1995).
24. Leturcq, D. J. *et al.* Antibodies against CD14 protect primates from endotoxin-induced shock. *J. Clin. Invest.* **98**, 1533–1538 (1996).
25. Ademar, A. & Ulevitch, R. J. Toll-like receptors in the induction of the innate immune response. *Nature* **406**, 782–787 (2000).
26. Vasselon, T. & Detmers, P. A. Toll receptors: a central element in innate immune responses. *Infect. Immun.* **70**, 1033 (2002).
27. Medzhitov, R., Preston-Hurlburt, P. & Janeway, C. A. Jr A human homologue of the *Drosophila* Toll protein signals activation of adaptive immunity. *Nature* **388**, 394–397 (1997).
28. Poltorak, A. *et al.* Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in *Tlr4* gene. *Science* **282**, 2085–2088 (1998).
29. Opal, S. M. & Huber, Ch. Bench-to-bedside review: Toll-like receptors and their role in septic shock. *Crit. Care* **6**, 125–136 (2002).
30. Takeuchi, O. *et al.* Differential roles of TLR2 and TLR4 in recognition of gram-negative and gram-positive bacterial cell wall components. *Immunology* **11**, 443–451 (1999).
31. Hayashi, F. *et al.* The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature* **410**, 1099–1103 (2001).
32. Shimazu, R. *et al.* MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. *J. Exp. Med.* **189**, 1777–1782 (1999).
33. Nagai, Y. *et al.* Essential role of MD-2 in LPS responsiveness and TLR4 distribution. *Nature Immunol.* **3**, 667–672 (2002).
34. Ozinsky, A. *et al.* The repertoire for pattern recognition of pathogens by the innate immune system is defined by cooperation between toll-like receptors. *Proc. Natl Acad. Sci. USA* **97**, 13766–13771 (2000).
35. Lorenz, E., Mira, J. P., Cornish, K. L., Arbour, N. C. & Schwartz, D. A. A novel polymorphism in the Toll-like receptor 2 gene and its potential association with staphylococcal infection. *Infect. Immun.* **68**, 6398–6401 (2000).
36. Lorenz, E., Mira, J. P., Frees, K. L. & Schwartz, D. A. Relevance of mutations in the TLR4 receptor in patients with gram-negative septic shock. *Arch. Intern. Med.* **162**, 1028–1032 (2002).
37. Bowie, A. & O'Neill, L. A. The interleukin-1 receptor/Toll-like receptor superfamily: signal generators for pro-inflammatory interleukins and microbial products. *J. Leukoc. Biol.* **67**, 508–514 (2000).
38. Werner, T. *et al.* A family of peptidoglycan recognition proteins in the fruit fly *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA* **97**, 13772–13777 (2000).
39. Liu, C., Xu, Z., Gupta, D. & Dziarski, R. Peptidoglycan recognition proteins: a novel family of four human innate immunity pattern recognition molecules. *J. Biol. Chem.* **276**, 34686–34694 (2001).
40. Michel, T., Reichhart, J. M., Hoffmann, J. A. & Royet, J. *Drosophila* Toll is activated by Gram-positive bacteria through a circulating peptidoglycan recognition protein. *Nature* **414**, 756–759 (2001).
41. Gottar, M. *et al.* The *Drosophila* immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. *Nature* **416**, 640–644 (2002).
42. Ramet, M., Manfrulli, P., Pearson, A., Mathey-Prevot, B. & Ezekowitz, R. A. Functional genomic analysis of phagocytosis and identification of a *Drosophila* receptor for *E. coli*. *Nature* **416**, 644–648 (2002).

43. Choe, K. M., Werner, T., Stoven, S., Hultmark, D. & Anderson, K. V. Requirement for a peptidoglycan recognition protein (PGRP) in Relish activation and antibacterial immune responses in *Drosophila*. *Science* **296**, 359–362 (2002).
44. Bouchon, A., Dietrich, J. & Colonna, M. Inflammatory responses can be triggered by TREM-1, a novel receptor expressed on neutrophils and monocytes. *J. Immunol.* **164**, 4991–4995 (2000).
45. Bouchon, A., Facchetti, F., Weigand, M. A. & Colonna, M. TREM-1 amplifies inflammation and is a crucial mediator of septic shock. *Nature* **410**, 1103–1107 (2001).
46. Inohara, N., Ogura, Y., Chen, F. F., Muto, A. & Nunez, G. Human Nod1 confers responsiveness to bacterial lipopolysaccharides. *J. Biol. Chem.* **276**, 2551–2554 (2001).
47. Inohara, N., Ogura, Y. & Nunez, G. Nods: a family of cytosolic proteins that regulate the host response to pathogens. *Curr. Opin. Microbiol.* **5**, 76–80 (2002).
48. Hampe, J. et al. Association of NOD2 (CARD 15) genotype with clinical course of Crohn's disease: a cohort study. *Lancet* **359**, 1661–1665 (2002).
49. Dinarello, C. A. Proinflammatory and anti-inflammatory cytokines as mediators in the pathogenesis of septic shock. *Chest* **112**, 321S–329S (1997).
50. Cohen, J. Adjunctive therapy in sepsis: a critical analysis of the clinical trial programme. *Br. Med. Bull.* **55**, 212–226 (1999).
51. Yang, H., Wang, H. & Tracey, K. J. HMG-1 rediscovered as a cytokine. *Shock* **15**, 247–253 (2001).
52. Wang, H. et al. HMG-1 as a late mediator of endotoxin lethality in mice. *Science* **285**, 248–251 (1999).
53. Bozza, M. et al. Targeted disruption of migration inhibitory factor gene reveals its critical role in sepsis. *J. Exp. Med.* **189**, 341–346 (1999).
54. Calandra, T. et al. Protection from septic shock by neutralization of macrophage migration inhibitory factor. *Nature Med.* **6**, 164–170 (2000).
55. Calandra, T., Spiegel, L. A., Metz, C. N. & Bucala, R. Macrophage migration inhibitory factor is a critical mediator of the activation of immune cells by exotoxins of Gram-positive bacteria. *Proc. Natl Acad. Sci. USA* **95**, 11383–11388 (1998).
56. Calandra, T. et al. MIF as a glucocorticoid-induced modulator of cytokine production. *Nature* **377**, 68–71 (1995).
57. Bosio, D. et al. Stimulation of toll-like receptor 4 expression in human mononuclear phagocytes by interferon-gamma: a molecular basis for priming and synergism with bacterial lipopolysaccharide. *Blood* **99**, 3427–3431 (2002).
58. Levi, M. & ten Cate, H. Disseminated intravascular coagulation. *N. Engl. J. Med.* **341**, 586–592 (1999).
59. Van der Poll, T., de Jonge, E., Levi, M. & van Deventer, S. J. Pathogenesis of DIC in sepsis. *Sepsis* **3**, 103–110 (1999).
60. Okajima, K. Regulation of inflammatory responses by natural anticoagulants. *Immunol. Rev.* **184**, 258–274 (2002).
61. Riewald, M., Petrovan, R. J., Donner, A., Mueller, B. M. & Ruf, W. Activation of endothelial cell protease activated receptor 1 by the protein C pathway. *Science* **296**, 1880–1882 (2002).
62. Faust, S. N. et al. Dysfunction of endothelial protein C activation in severe meningococcal sepsis. *N. Engl. J. Med.* **345**, 408–416 (2001).
63. Hotchkiss, R. S. et al. Apoptotic cell death in patients with sepsis, shock, and multiple organ dysfunction. *Crit. Care Med.* **27**, 1230–1251 (1999).
64. Hotchkiss, R. S. et al. Sepsis-induced apoptosis causes progressive profound depletion of B and CD4⁺ T lymphocytes in humans. *J. Immunol.* **166**, 6952–6963 (2001).
65. Hotchkiss, R. S. et al. Prevention of lymphocyte cell death in sepsis improves survival in mice. *Proc. Natl Acad. Sci. USA* **96**, 14541–14546 (1999).
66. Döcke, W.-D. et al. Monocyte deactivation in septic patients: restoration by IFN- γ treatment. *Nature Med.* **3**, 678–681 (1997).
67. Stuber, F. Effects of genomic polymorphisms on the course of sepsis: is there a concept for gene therapy? *J. Am. Soc. Nephrol.* **12**(Suppl. 17), S60–S64 (2001).
68. Landry, D. W. & Oliver, J. A. The pathogenesis of vasodilatory shock. *N. Engl. J. Med.* **345**, 588–595 (2001).
69. Brealey, D. et al. Association between mitochondrial dysfunction and severity and outcome of septic shock. *Lancet* **360**, 219–223 (2002).
70. Rivers, E. et al. Early goal-directed therapy in the treatment of severe sepsis and septic shock. *N. Engl. J. Med.* **345**, 1368–1377 (2001).
71. Van den Berghe, G. et al. Intensive insulin therapy in critically ill patients. *N. Engl. J. Med.* **345**, 1359–1367 (2001).
72. Annane, D. et al. Effects of treatment with low doses of hydrocortisone and fludrocortisone on mortality in patients with septic shock. *J. Am. Med. Assoc.* **288**, 862–871 (2002).
73. Bernard, G. R. et al. Efficacy and safety of recombinant human activated protein C for severe sepsis. *N. Engl. J. Med.* **344**, 699–709 (2001).
74. Bochkov, V. N. et al. Protective role of phospholipid oxidation products in endotoxin-induced tissue damage. *Nature* **419**, 77–81 (2002).

Contacts

Publisher: Ben Crowe

Editor: Paul Smaglik

Marketing Manager: David Bowen

European Head Office, London

The Macmillan Building
4 Crinan Street
London N1 9XW, UK
Tel +44 (0) 20 7843 4961
Fax +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Senior European Sales Manager:
Nevin Bayoumi (4978)

UK/ RoW/ Ireland:

Matt Powell (4953)

Andy Douglas (4975)

Frank Phelan (4944)

Netherlands/ Italy/ Iberia:

Evelina Rubio Hakansson (4973)

Scandinavia: Sille Opstrup (4994)

France/ Belgium:

Amelie Pequignot (4974)

Production Manager: Billie Franklin

To send materials use London
address above.

Tel +44 (0) 20 7843 4814

Fax +44 (0) 20 7843 4996

e-mail: naturejobs@nature.com

International

Advertising Coordinator:

Hind Berrada (4935)

Naturejobs web development:

Tom Hancock

Naturejobs online production:

Ben Lund

European Satellite Office

Germany/ Austria/ Switzerland:

Patrick Phelan, Odo Wulffen

Tel + 49 89 54 90 57 11/-2

Fax + 49 89 54 90 57 20

e-mail: p.phelan@nature.com

o.wulffen@nature.com

US Head Office, New York

345 Park Avenue South,
10th Floor, New York, NY 10010-1707

Tel +1 800 989 7718

Fax +1 800 989 7103

e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

US Advertising Coordinator:

Linda Adam

Japan Head Office, Tokyo

MG Ichigaya Building (5F),

19-1 Harakakamachi,

Shinjuku-ku,

Tokyo 162-0841

Tel +81 3 3267 8751

Fax +81 3 3267 8746

Asia-Pacific Sales Director:

Hideki Watanabe

e-mail: h.watanabe@naturejpn.com

naturejobs

Converging on DNA

In next year's calendar of scientific events, occasions marking the fiftieth anniversary of the elucidation of DNA's twisted structure span continents, disciplines and dress codes. This *mélange* of events, listed in the directory that follows, is somehow appropriate when you consider the backgrounds and interests of the principals involved in the discovery.

Rosalind Franklin was a British biochemist who, according to a recent biography, was more comfortable living in France than in England. She worked at the University of Cambridge with Maurice Wilkins, a physicist born in New Zealand, using X-ray crystallography to discern DNA's structure.

They alternately competed and cooperated with another team arising from equally mixed disciplines and origins. James Watson was a biologist from the United States who was working with Francis Crick, an English physicist. In their Cambridge lab, the two used their knowledge of chemical bonds, as well as X-rays of DNA generated by Franklin and Wilkins, to solve the molecule's structure.

It is fitting, then, that events celebrating this finding are all over the map; literally and figuratively. They include genetic symposia in France, artistic exhibits in England and at least one black-tie dinner in New York. It is also appropriate that DNA-related meetings are scattered beyond places that have obvious geographical connections to the discovery of DNA's double helix. Contributions by scientists across the globe, from all disciplines, to the application of this ground-breaking work transcend — but don't overshadow — the finding that emerged from two English labs.

Everyone, understandably, wants a piece of the double helix. This year's calendar of events will help to make that possible.

Paul Smaglik

Naturejobs editor



Contents

EVENTS DIRECTORY 2003

WWW.NATUREJOBS.COM

Career centre
Information on the
scientific job market

FOCUS

SPOTLIGHT

RECRUITMENT

SCIENTIFIC ANNOUNCEMENTS

SCIENTIFIC EVENTS